# PULSES AND TRANSIENTS IN COMMUNICATION CIRCUITS

# PULSES AND TRANSIENTS IN COMMUNICATION CIRCUITS

*An Introduction to Network Transient Analysis for Television and Radar Engineers*

BY

COLIN CHERRY
M.Sc. (Eng.), A.M.I.E.E.

*Lecturer to Imperial College*
*University of London*

*Late of the Research Laboratories of the*
*General Electric Co., Ltd., Wembley*

LONDON
CHAPMAN & HALL LTD
37 ESSEX STREET, W.C.2

1949

*First published 1949*

THIS BOOK IS PRODUCED IN COMPLETE CONFORMITY WITH THE AUTHORIZED ECONOMY STANDARDS

*Catalogue No. 361/4*

PRINTED AND BOUND IN ENGLAND BY WILLIAM CLOWES AND SONS LIMITED
LONDON AND BECCLES

"But the real desideratum, which, if it can be reached, is of paramount importance, is to get solutions which can be understood and appreciated at first sight and followed into detail with ease, presenting to us as nearly as possible the effects as they really occur in the physical problem, disconnected from the often unavoidable complication due to the form of mathematical expression."

OLIVER HEAVISIDE

*Electrical Papers*. Vol. II, p. 308.

# PREFACE

This book is intended as an introduction to circuit transient analysis, for communications engineers, which will bridge the gap between simple conventional alternating current theory and operational methods of analysis. A great deal of excellent material has been published during recent years, concerning the transient behaviour of communication circuits; much of this has been written in the form and style of the professional mathematician and some has been applied directly to engineering systems such as television, pulse modulation, radar and the like. Nevertheless, the basis of the subject must be mathematical and this provides difficulties for many an engineer, to whom mathematics may be only one aspect with which he must be acquainted in his many-sided profession.

The great difficulty which many people experience when setting out to supplement their knowledge of steady-state circuit theory by a study of transients (and this is becoming increasingly necessary) is to know where to start. This book is then intended to give such readers the essential groundwork, using, wherever possible, rigorous physical arguments and only elementary mathematics. References to many published books and papers are given at suitable points in the text in order to lead the reader up to this great mass of literature. Mathematical language and notation are avoided as much as possible; electric *waveforms* are dealt with, rather than analytical functions, thus giving the book a geometrical (or "oscillographic") flavour.

Chapter 1 is very elementary, being concerned with establishing the physical ideas behind the solution of the differential equations of the type which arise with linear circuits and the essentially exponential shape of all electric waveforms. Further, that old mystery is discussed: the "Particular Integral" and the "Complementary Function"—the need for two solutions. The waveforms and spectra of electric signals are examined in more detail in Chapter 2, in the light of Fourier analysis; the great simplicity of the conjugate vector system is emphasised, which appears to be somewhat neglected in elementary texts. This same system of conjugates is employed in the next Chapter as a means of describing the selectivity characteristics of communication networks, some of the main physical and geometrical

properties of which are closely examined. The limitations which these properties impose on the propagation of waves and transient signals through the networks form the basis of Chapter 4, while the succeeding Chapter deals with the various means of simplifying transient response calculations which result from the use of "idealised characteristics". Much erroneous and misleading information has been published concerning such non-physical characteristics, but their importance, in view of the simple mathematics they offer, should not be underestimated. Care should be taken not to draw false conclusions, by too exact interpretation of their results; their use and abuse is discussed here.

Chapter 6 is concerned with chains of amplifier stages in a communication channel and with their transient responses, particular attention being paid to the accuracy of signal reproduction, a matter of great importance in the design of many wide-band amplifiers, such as those used for television. The peculiar effects resulting from the unequal distortion of the two sets of sideband of a modulated carrier (as in single- or asymmetric sideband channels) has mostly been treated only in approximate ways in the literature and it is hoped that Chapter 7 remedies this. In that Chapter a simple method is described by which means the distortion of either steady-state or transient signals may be calculated, with any depth of carrier modulation and any type of channel characteristics.

The last Chapter is primarily concerned with the response of networks to very short impulses. This, besides giving a neat physical interpretation of signal distortion by the idea of echoes, provides a very clear picture of Fourier Transforms, by virtue of the essentially inverse nature of *frequency* and *time* as variables, at the same time avoiding most of the complexities of mathematical expression.

In conclusion, I should like to express my indebtedness to the late Sir Clifford Paterson and to members of his staff at the Research Laboratories of the General Electric Co. Ltd. (Wembley, England) for much help and encouragement in writing this book. Also I should like to acknowledge the kind permission which the Institution of Electrical Engineers (London) has given me, to make use of material and drawings from my papers already published in their Journal. For reading and checking the final proofs, I am most grateful to A. R. Boothroyd.

COLIN CHERRY

MANCHESTER,
*October*, 1948.

## CONTENTS

# SYMBOLS

| | |
|---|---|
| $ab$, $cd$, $ef$ | indicates network's terminals. |
| $a_0$, $a_n$, $b_n$ | amplitudes of cosine and sine harmonics. |
| $a(\omega)$, $b(\omega)$ | cosine and sine component amplitudes of spectrum. |
| $A+jB\ (=\gamma)$ | propagation constant (attenuation and phase components). |
| $\alpha$, $\alpha_c$ | a "universal" frequency scale. |
| $\alpha_n=(a_n+jb_n)$ | infinitesimal component amplitude. |
| $B$ | band-width. |
| $\beta$ | $=\sqrt{(R^2/4L^2-1/LC)}$. |
| $C$, $S$ | cosine and sine components, of wave or network. |
| $C_v=C_{ac}+C_{gc}$ | the total valve capacity. |
| $e$, $e(t)$ | E.M.F. (instantaneous). |
| $e_E$ | envelope voltage wave. |
| $\epsilon$ | natural logarithmic base. |
| $\eta$ | ratio of sideband/carrier amplitude. |
| $f$ | frequency. |
| $f(t)$ | a waveform; $f_c(t)$, $f_s(t)$, odd and even components. |
| $G$ | overall amplifier gain at zero (or mid-band) frequency. |
| $\Delta G$ | a specific drop in gain. |
| $g$ | valve mutual conductance. |
| $\gamma$ | propagation constant; $R$, $L$, $G$, and $C$ primary constants. |
| $h(t)$ | response of circuit to unit step wave. |
| $I$, $V$ | peak current and voltage. |
| $i$, $i(t)$ | current (instantaneous). |
| $I$ | ideally narrow pulse. |
| $I_E$ | varying carrier amplitude. |
| $I_0$ | D.C. or zero frequency component; also steady carrier amplitude; also initial current. |
| $j$ | an operator ; $j^2=-1$. |
| $K$ | $=T_0/T_1$. |
| $k$ | deviation ratio; also used for $L/CR^2$. |
| $k'$ | phase deviation ratio. |
| $\kappa$ | special value of $\omega CR$. |
| $L$, $C$, $R$ | circuit elements. |
| $l$ | length of line or cable. |
| $m$ | a constant; also for depth of modulation. |
| $\mu$ | valve amplification factor. |

| | |
|---|---|
| $N$ | specific whole number; number of circuit elements or valve stages. |
| $n$ | harmonic number. |
| $OA$, $OB$ | vector lengths (on diagrams). |
| $\Omega$ | $\omega CR$. |
| $\omega$ | $=2\pi f$. |
| $\omega_1$ | idealised cut-off frequency. |
| $\omega_m$ | frequency of modulating wave component. |
| $\omega_o$ | steady sine wave (carrier) frequency. |
| $\omega_s$ | a specific frequency. |
| $\hat{\omega}$ | instantaneous frequency. |
| $p$ | $=j\omega$. |
| $\phi$, $\theta$ | phase angles. |
| Q | circuit "goodness." |
| $Q_0$ | initial charge. |
| $r$ | specific harmonic. |
| $R_T$, $R_1$, $R_2$ . . ., etc. | terminating resistors. |
| $ROR'$ | real axis. |
| $SOS'$ | imaginary axis. |
| $T$ | transfer ratio. |
| $T_0$ | repetition period. |
| $T_1$ | pulse duration. |
| $t$ | time. |
| $t_1$ | time delay. |
| $\tau$ | a specified time, a value of $t$. |
| $v$, $v(t)$ | voltage (instantaneous), usually a *response*. |
| $w$ | a frequency range. |
| $X$ | reactance. |
| $x$ | geometric distance. |
| $[X]$ and $[Y]$ | single sideband, in-phase and quadrature components. |
| $Y_x(\omega)$, $Y_y(\omega)$ | single sideband, in-phase and quadrature admittances. |
| $g_x(\omega)$, $b_x(\omega)$, $g_y(\omega)$, $b_y(\omega)$ | single sideband, in-phase and quadrature admittances. |
| $Y_l$, $Y_u$ | admittances to lower and upper sidebands. |
| $Z$, $z(\omega)$ | impedance. |

# CHAPTER 1

## INTRODUCTORY

## THE BASIS OF NETWORK ANALYSIS

### 1. Introduction

Interest in the transient response of networks, particularly communication networks, is increasing amongst engineers; this is partly due to its importance in radar work, which has engaged the attention of many during the last few years, partly due to work on television, picture telegraphy, and high-speed telegraphy, and partly due to the increased attention which television is expected to receive during the post-war years. These subjects, and others, are ones in which the engineer continually comes up against the problem of signal *waveform* distortion; it is not sufficient to state a certain bandwidth or selectivity, but the response to waves of given shape must be known—pulses, or steep-edged waves, signals expressive of the "definition" of a channel characteristic and of the ability of a channel to pass and reproduce a signal at its output terminals as nearly as possible of the same *waveform* shape as that of the input signal.

The fact that so many engineers are now concerned with transient behaviour of networks does not mean that the subject is one which has arisen only during recent years; it is really a revival of interest, since some of the earliest work done on communication circuits concerned their response to suddenly applied E.M.Fs.[1]—and the waveshape aspect was the important one. This was natural since the laws of network behaviour to a suddenly applied wave were derived from the differential equations of circuit meshes, which involved currents and voltage as functions of *time*. The working out of these equations led to the ideas of "transient" solutions and "steady-state" solutions, the former giving the waveform of the current or voltage response during the initial shock of the applied wave and the latter the response after some period of time when the disturbance has settled down.

Study of the steady-state solutions led to development of the familiar alternating-current theory and the use of A.C. vectors.[2]

It is partly due to the extensive teaching of A.C. theory, and its comparative mathematical simplicity, and partly to the extensive manufacture and use of measuring instruments and other apparatus involving sine-wave techniques that the average engineer has, so to speak, a sine-wave background to his ideas and practices. The sine-wave A.C. vector theory is familiar ground to him, and in many cases to venture into the field of transient analysis may be a confusing experience. To many the subject seems a conglomeration of ideas involving difficult operational methods and notations and seems one far more suitable for attack by the professional mathematician; frequently, too, an approximate solution only is required, giving an idea of what the general character of the network response will be, and it may seem hardly worth while to involve the relative difficulties of such mathematical methods.

We shall not be concerned primarily with rigorous proofs of mathematical methods in this book, but it is hoped that the material in here will provide a basis for understanding the significance of these methods for those wishing to study them.[3, 12, 13, 14, 15] We shall be emphasising the geometrical aspect here rather than the purely analytical, that is to say the waveform *shapes* will continually be referred to rather than analytical expressions for them. An attempt will be made to classify the various types of wave and network characteristics in such a way that the mechanism of transient response is brought out; starting from the steady-state response to sine-waves, the relation between the sine-wave and transient behaviour of networks will be studied, using methods which will be familiar to engineers.

Before starting on the main body of the book, we shall consider a few fundamental aspects of the response of a network to an applied wave, to clear the ground and to establish some definitions. Perhaps the matter in this introductory chapter is already familiar to the reader, but it is intended only as a brief review of some of the basic principles.

## 2. Fundamental laws of network behaviour

However complex some methods of transient response analysis may appear, they are all dependent on the fact that currents and voltages in linear networks may be represented by linear differential equations with constant coefficients. A definition of "linear network" will be given in the next paragraph. It is assumed that the student is familiar with such equations to some extent,[6] since this

chapter is intended to be a short review of some of the physical principles involved, and expressed in mathematical shorthand by such equations.

These differential equations are formed from the laws for voltages and current flow in the three kinds of element that we have in electric circuits—inductance, capacity, and resistance ($L$, $C$, and $R$) —the first two capable only of storing magnetic and electrostatic energy and the last capable only of dissipating it as heat. The laws for relating the voltages and currents, $v$ and $i$, are of course:*

$$v=L\frac{di}{dt}, \qquad i=C\frac{dv}{dt}, \qquad v=Ri \quad . \; . \; . \quad (1)$$

whatever the waveform of $v$ or $i$; or, correspondingly:

$$i=\frac{1}{L}\int vdt, \qquad v=\frac{1}{C}\int idt, \qquad i=\frac{1}{R}v \quad . \; . \quad (2)$$

Now by "linear network" we mean one in which the elements $L$, $C$, and $R$ are constants, independent of current in them or voltage across them. Thus the geometric shape of a current waveform in such a network will be independent of the magnitude of the voltage producing it, but will depend only on the shape of the voltage waveform. We shall be concerned entirely with linear networks throughout this book.

By the use of Kirchhoff's laws† we may add up the voltages round any *mesh* (i.e. a complete, closed, electrical path), or alternatively the currents entering any *branch-point* in the network, and equate them to zero, to form a differential equation for the mesh or for the circuit branch-point. Fig. 1 will clarify this; in (a) a voltage generator having any waveform $e(t)$ of E.M.F. drives a current through the elements $R$, $L$, and $C$ in series, and in (b) a current generator of waveform $i(t)$ supplies the three elements $R$, $L$, and $C$ in parallel, and the voltage drop $v(t)$ appears across

* Assuming, as is usual, that the circuit elements $L$, $C$, and $R$ do not vary with time. The laws were originally framed in terms of rate of change of flux linkages, $d(Li)/dt$ and $d(Cv)/dt$.

† Kirchhoff's laws:

(1) The algebraic sum of the E.M.Fs. and potential drops round any closed path (mesh) must be zero.
(2) The total current entering or leaving any point in a circuit must be zero.

These laws apply to every instant of time, and hence are independent of current or voltage waveforms.

them between the points $AB$. We use $e(t)$ and $v(t)$ as symbols to distinguish between E.M.F. and voltage drop.

Adding up the voltages around the mesh in (a) we obtain:

$$L\frac{di}{dt}+Ri+\frac{1}{C}\int idt-e=0 \quad . \quad . \quad . \quad . \quad . \quad (3)$$

[omitting the qualifying symbol ($t$) for simplicity, but still assuming that $i$ has the waveform $i(t)$].

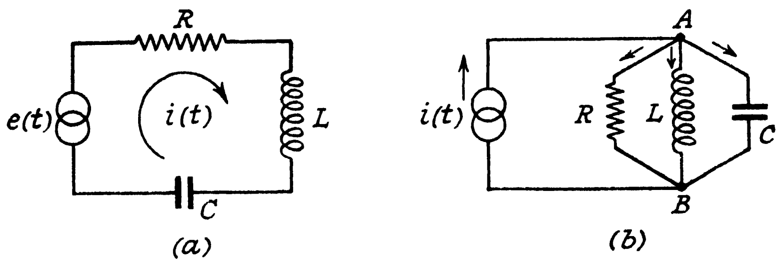



Fig. 1.—(a) Mesh and (b) Branch Currents.

Similarly in (b), adding up the currents entering the point $A$ or point $B$, we have:

$$C\frac{dv}{dt}+\frac{1}{R}v+\frac{1}{L}\int vdt-i=0 \quad . \quad . \quad . \quad . \quad . \quad (4)$$

[again writing $v$ instead of $v(t)$ for simplicity].

These equations then give the relation between the waveform shapes of $e$ and $i$ and between $v$ and $i$ in these two elementary circuits. These circuits could be considered as being parts of more

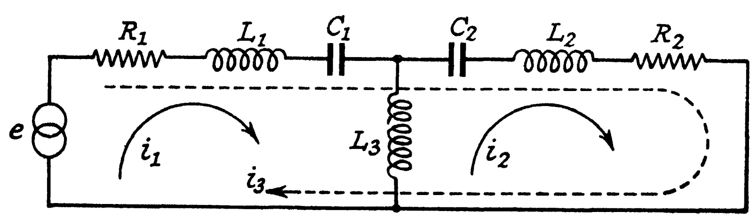



Fig. 2.—A Two-mesh Network, showing a "Redundant" Mesh Current (dotted line).

complicated structures, for which we should need a number of such equations to determine the waveforms at any point. For example, Fig. 2 shows a two-mesh network. In order to determine the current or voltage waveforms anywhere in this network, given the

waveform of $e$ [again an abbreviation for $e(t)$], we could form two differential equations for the two meshes in which the currents $i_1$ and $i_2$ are shown flowing, thus:

$$L_1\frac{di_1}{dt}+L_3\frac{d}{dt}(i_1-i_2)+R_1i_1+\frac{1}{C_1}\int i_1dt-e=0 \quad . \quad . \quad . \quad (5)$$

and $$L_2\frac{di_2}{dt}+L_3\frac{d}{dt}(i_2-i_1)+R_2i_2+\frac{1}{C_2}\int i_2dt=0 \quad . \quad . \quad . \quad (6)$$

We see there is no generator E.M.F. term in the second mesh, equation 6, but the place of the generator as the means of supplying the energy to the mesh is taken by the voltage appearing across $L_3$, that is $L_3\frac{d}{dt}(i_2-i_1)$. This common element $L_3$ is hence known as the *coupling impedance* between the two meshes.

We could equally well have obtained a third equation by taking the sum of the voltages round the third "closed path" as indicated by the dotted arrow on the circuit of Fig. 2, but this would have been an unnecessary one, since the current that it would have expressed is already known in terms of $i_1$ and $i_2$. Such a mesh is known as a *redundant mesh*, while the least number of meshes required to define the current in every single element in the circuit gives the number of *independent meshes.*

Each mesh in any linear network may contain a number of the three kinds of element, inductance, capacity, and resistance, and the mesh equations will be of the general type 5 or 6 in that they will represent the sum of the voltage waveforms appearing across inductance, capacity, and resistance elements which, at every instant of time, will be identical with the sum of the waveforms of the applied E.M.Fs. in the mesh. Then the solution of equations of this general type will give considerable information concerning the relation between voltage and current waveforms in networks, that is, the transient behaviour. For an elementary discussion such as this it is sufficient to take the simplest kind of mesh equation (equation 3), repeated here:

$$L\frac{di}{dt}+Ri+\frac{1}{C}\int idt=e \quad . \quad . \quad . \quad . \quad . \quad . \quad [(3)]$$

representing the behaviour of the mesh in Fig. 1 (a).

In this introductory chapter we refer continually to the *energy* stored in the $L$ and $C$ elements in this simple mesh, so it is of interest

to rewrite equation 3 as an equation of energy. Multiplying both sides by $i$ gives:

$$\frac{d}{dt}(\tfrac{1}{2}Li^2)+Ri^2+\frac{d}{dt}\left(\tfrac{1}{2}\frac{q^2}{C}\right)=ei \quad . \quad . \quad . \quad . \quad (7)$$

where $q=$the instantaneous charge in the condenser$=\int_o^t idt$. On the right-hand side we have the power or rate at which energy is being supplied to the mesh from an external source, while on the left we have the rate of dissipation of energy in the resistance plus the rate at which energy is changing in the inductance and capacity. These energies are specified, at any instant of time, by the current $i$ in the inductance and the charge $q$ in the condenser.

If one of the "storage" elements $L$ or $C$ were omitted we should

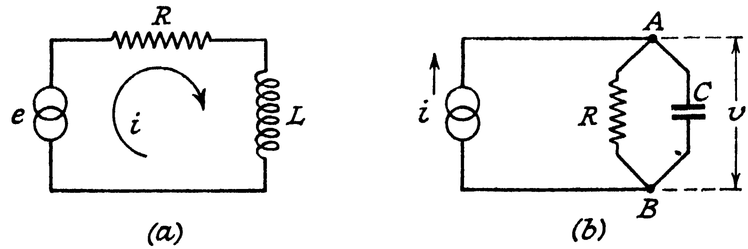



Fig. 3.—Elementary Circuits with one Energy-storing Element.

have (for instance omitting $C$, giving the circuit in Fig. 3 (a)) an equation with one less term:

$$L\frac{di}{dt}+Ri=e \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8a)$$

Similarly, in equation 4 representing the behaviour of the circuit in Fig. 1 (b), we should have, if $L$ were omitted as in Fig. 3 (b):

$$C\frac{dv}{dt}+\frac{1}{R}v=i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8b)$$

These equations may be regarded as special cases of the general equations 3 or 4.

There is an important relation between the currents given by equations 3 and 8a and similarly between equations 4 and 8b, which we shall show later* leads to a very useful theorem concerning

* The low-pass/band-pass equivalent network theorem—see Sec. 33.

"low-pass" and "band-pass" filters with similarity in their transient response characteristics. Now both equations 3 and 8a represent a current flowing round a single mesh under the driving force of an applied E.M.F., $e$, and a circuit of this type is said to possess one *degree of freedom*, since only one current can flow, i.e. the current round the one mesh. A network possessing $N$ meshes, without considering redundant meshes, may have all its constituent mesh currents represented by a series of $N$ differential equations. Such a network has $N$ degrees of freedom, since $N$ independent currents flow when the E.M.F. is applied; by "independent" we mean that no one mesh current can be expressed in terms of the others. These meshes may each contain two types of energy-storing element, $L$ and $C$, resulting in an equation of the form 3, or one type only, requiring an equation of the form 8b; this does not affect the number of degrees of freedom, but only the type of waveform of the mesh current, as we shall see later.

Before proceeding with a discussion of the physical aspects of these basic differential equations it would be well to emphasise the similarity between the two equations 3 and 4, which give the relations between the waveforms of the currents and voltages in the simple circuits of Fig. 1 (a) and (b). These equations are similar, term by term, in that $L$ in one equation is replaced by $C$ in the other, $R$ replaced by $1/R$ and the voltage generator E.M.F., $e$, replaced by the current generator supplying $i$ (we have used different voltage symbols $e$ and $v$ merely to distinguish between a generated E.M.F. and a potential drop). Regarded purely as a mathematical problem we should expect these equations to lead to identical solutions, one giving the waveform of $i$ and the other the waveform of $v$.

Networks such as those in the figure are said to be *dual networks*, and the differential equations representing their behaviour are similar in the way we have described. Almost any network may be shown to have a dual, however many independent meshes it may contain, in which case the two networks will be represented (as regards their response) by sets of corresponding differential equations. The similarity should be noted between these elementary dual networks of Fig. 1 (a) and (b) with regard to the nature and positions of the elements which correspond in the equations; a condenser in series with the generator corresponds to an inductance in shunt with the generator, and so on. This theorem is extremely useful, since any transient response calculation we may make on one network may

immediately be applied to the other. For instance, in these simple circuits if we know the waveform of the applied E.M.F., $e$ (in (a)), we may calculate the corresponding current waveform, $i$, by equation 3; then exactly the same relation would hold between the waveforms of the generator current $i$ and the resulting voltage $v$ in (b). Only one equation need be solved for both circuits. We shall be dealing with some more practical examples later in Chapter 3, Sec. 28, which will illustrate the use of this important theorem, giving the rules for forming the dual of a network and showing when it is not possible.

As a matter of interest it may be mentioned that the use of certain mechanical analogies to electric circuits is only justified because the differential equations representing their behaviour are similar to those described above for electric meshes. Relatively complicated mechanical systems may be dealt with in a manner very much like the division of an electric network into its independent meshes and their behaviour expressed by a series of equations of the basic type described. Mechanical systems were very much studied in the early days of communications, particularly vibrating strings, and the results applied to electric systems on the basis of the similarity of the equations. Some references are given for those readers interested in the subject of mechanical analogies to electric systems.[7, 8, 9, 17, 18]

## 3. Physical meaning of the "solutions" of network equations

In this section we shall consider what is meant by the "solution" of the basic mesh equations 3 or 8a in a physical way. The reader may already be familiar with the mathematical method of dealing with such equations, but in this book we are not concerned so much with pure analytical methods as such, but rather with the geometry of current and voltage waveforms, and we shall consider the present problem from this point of view.

In the circuit of Fig. 1 (a) we have an applied E.M.F. which has been written there as $e(t)$, meaning that it is a function of time—which is what is meant by a "waveform." We require to know the corresponding current (which is flowing by virtue of this applied E.M.F.) as a function of time, that is, to know the current waveform. The relations between the current and the applied E.M.F. is, of course, given by equation 3, but we cannot directly see the form of the current from this, since we cannot readily visualise differentials and integrals in our minds in a general way. Mathematically

speaking, we say that the current function is implicit in the equation, meaning that its form is implied or contained in the equation but not stated in such a way as to be understood directly, whereas we know from experiment that current and voltage waveforms in electric circuits are explicit functions of time—they have definite magnitudes at every instant of time. We know, from physical reasoning, that the current waveform must be of finite amplitude, single valued (i.e. current cannot have two values at once) and continuous, if we are to be practical and ignore the possibility of infinite voltages ; these three conditions must also apply to the three waveforms of the voltages across the inductance, resistance, and capacity elements as represented by the first three terms in the equation 3. We shall show later that the applied driving E.M.F. (or current) may theoretically be discontinuous, since it may be produced by switching a generator of some kind, but in practice the result of this applied driving force must obey these conditions, since every physical circuit must possess a finite amount of inductance and capacity although they may not be shown on a diagram. We are putting these particular points forward here only in an endeavour to find out what general type or class of mathematical function must be used to express the shape of the current waveform and to show the physical limitations of such an expression. Once having settled on this "solution" that satisfies the physical conditions, then we may leave it to the mathematics to see what the effects are of making certain idealisations—for instance that of using currents and voltages having discontinuities by ignoring residual inductance or capacity in a circuit.

To return now to equation 3, we should expect the function giving the current waveform to separate out of the equation as a discrete term, since we know that both the applied E.M.F. and the current have definite values at every instant of time. Thus equation 3 must be expressible in the form:

(Current waveform) $\times$ (Network function)
$= e$, the applied driving E.M.F. waveform . . . (9)

We have referred to the second factor as a "network function" since clearly it relates the E.M.F. and current waveforms, and hence must depend on the elements in the network. No other conditions are stated as yet.

For equation 3 to take the form of equation 9 the current waveform must be a function whose differential and integral are

proportional to, or equal to, the function itself. The only function of which this is true is the exponential one:

$$i=I\,.\,\epsilon^{mt} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (10)$$

where $m$ is any constant, independent of time.

With this waveform function we can visualise the differential and integral at the same time, as mentioned earlier in this section, since they are identical with the current waveform itself. The sum of the voltages round the mesh, as given by equation 3 now becomes:

$$I\,.\,\epsilon^{mt}\,.\left[mL+R+\frac{1}{mC}\right]=e \quad . \quad . \quad . \quad . \qquad (11)$$

so that our network function is $[mL+R+1/mC]$.

The reader will see that this network function is really characteristic of the elements in the mesh and their layout and may be said to define the mesh. This function is obviously equal to the ratio $e/i$ (applied E.M.F./resulting current).

But our argument that the current waveform $i$ must be exponential seems to have led to the statement in equation 11 that $e$ can only be exponential, since the network function $[mL+R+1/mC]$ is a constant. The equation 11 would be perfectly correct if this were so and the applied E.M.F. would have to be of the form:

$$e=E\,.\,\epsilon^{mt} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (12)$$

where $\quad E=I[mL+R+1/mC]$

But our argument concerning equation 9 applies equally well to a linear sum of any number of exponentials—and so currents may be represented as a series of elementary currents all having exponential waveforms, so that the total current is equal to the sum of these, in this way :

$$i=I_1\epsilon^{m_1t}+I_2\epsilon^{m_2t}+I_3\epsilon^{m_3t}+\ .\ .\ .\ ,\text{ etc.} \quad . \quad . \qquad (13)$$

again without restrictions on the constants $m_1$, $m_2$, $m_3$ . . .

Any current flowing in a physical network must be capable of being split up into such a series of exponential *current components* in order that the form of the current may be made explicit. Hence the E.M.F. giving rise to this current is capable of being considered as the sum of a number of exponential waveform components of voltage; thus by substituting for $i$ in equation 3:

$$I_1\epsilon^{m_1t}\,.\left[m_1L+R+\frac{1}{m_1C}\right]+I_2\epsilon^{m_2t}\,.\left[m_2L+R+\frac{1}{m_2C}\right]+\ .\ .\ .\ ,\text{ etc.}$$
$$=e \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (14)$$

This notion of dividing a voltage and current wave into component waves gives the basis of certain methods of transient analysis and methods of attacking the problem of the distortion of a wave by its passage through electric filter and other transmission networks, a problem which arises so frequently in communication engineering. One of the most important methods of analysis depends on the use of an imaginary value of $m$ in the exponential term. Thus the component waves become sinusoidal in shape.*

$$i=I\,.\,\epsilon^{j\omega t}$$

or

$$i=I[\cos\omega t+j\sin\omega t] \quad . \quad . \quad . \quad . \quad . \quad (15)$$

where $j\omega=m$, $\omega=$constant, $j^2=-1$.

The mathematical difficulty of the $j$ appearing in this expression, which we have stated as representing a sinusoidal wave, need not concern us here, but is discussed in the next chapter. Anyway, it may be noted that the arguments concerning the exponential waveforms of currents would apply to the real part of the exponential in this way:

$$i=\text{Real part of } I\,.\,\epsilon^{j\omega t} \quad . \quad . \quad . \quad . \quad . \quad (16)$$
$$=I\cos\omega t$$

so that the continuous sinusoidal waveform is one possible form for our elementary current components.

It should be emphasised that no limit was placed on the number of current or voltage components required to represent a complete physical waveform, and in practice an infinite number of such components may be required. Another point of importance arises concerning the waveforms of these current components, in that the time variable $t$ must be taken to be continuous from $-\infty$ to $+\infty$ and the exponential function must exist for all times. It must not be confused with the exponential wave *applied at* $t=0$ (shown in Fig. 4 (b) or (c)) and having zero amplitude before $t=0$. There is a great distinction to be drawn between the response of a network to such a wave, which is suddenly applied at a certain time and the waveform of which involves a discontinuity, and the response to a wave of the type we have so far been considering, having a continuous waveform and defined by its mathematical expression for all values of time from $t=-\infty$ to $+\infty$. In the latter case the whole history of the network is defined, but in the former case the waveform of the applied E.M.F. is defined only for the time after $t=0$, and the mathematical process for determining the resulting current

* Refer to De Moivre's theorem in any elementary algebra book.

relies on a knowledge of the E.M.F. waveform after $t=0$, but a condition is used in the mathematics which ensures that the previous history of the network is taken into account.

Thus in this case, if some energy had been stored in the capacity and inductance before the time $t=0$, we should expect this energy to modify the form of the resulting current, since it gives rise to a current quite apart from whether the E.M.F. be suddenly applied at $t=0$ or not. But if our applied E.M.F. is continuous and defined for all times, and is the only E.M.F. applied, then there can be no doubt about stored energies. Then if our applied E.M.F. be analysed into its sinusoidal waveform components, the resulting current may be calculated. The method of analysing a wave into such components is discussed in Chapter 2 ; a sinusoidal shape is not the only form for elementary E.M.F. and current waves, but it

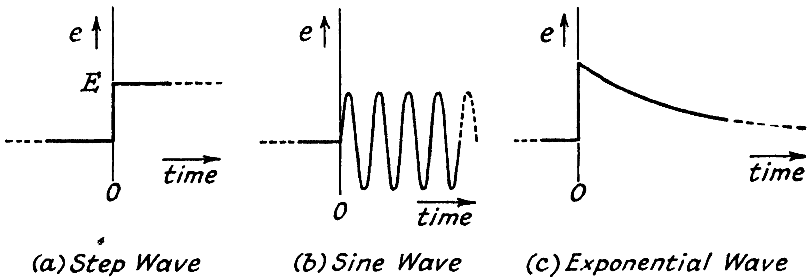



Fig. 4.—Discontinuous Driving E.M.F.'s, having Zero Amplitude before Time $t=0$.

is one which is extremely useful and commonly used, and, though mathematically limited, the Fourier method of circuit response analysis (which is based on such elementary wave components) is nevertheless very useful indeed for helping in the understanding of transient behaviour and of linear circuit distortion. It will be used as the background to the subject under consideration in this book.

## 4. Free and forced oscillations in a network

Many network problems concern the response to an applied wave having zero amplitude before a certain time $t=0$ and a definite amplitude after this time. Three typical examples of such waves are shown in Fig. 4. Here we have (a) a step wave, which may be defined as :

$$\left.\begin{array}{l} e=0,\ t<0 \\ e=E,\ t>0 \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (17)$$

Such a wave function is sometimes known as a Heaviside Unit Function,[10] and is then given the symbol $H$ or 1. In Fig. 4 (b) a step-modulated sine wave is shown, and may be defined as :

$$\left.\begin{array}{l} e=0,\ t<0 \\ e=E\cos(\omega_0 t+\phi),\ t>0 \end{array}\right\} \quad . \quad . \quad . \quad . \quad (18)$$

sometimes referred to as a "suddenly applied sine wave," since it may be regarded as a sinusoidal wave modulated in amplitude by the step wave of equation 17.

In Fig. 4 (c) a wave is shown which starts with an initial amplitude $E$, but which decays exponentially. Then in this case :

$$\left.\begin{array}{l} e=0,\ t<0 \\ e=E\,.\,\epsilon^{-mt},\ t>0 \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

Now let us consider the behaviour of the mesh shown in Fig. 1 (a) when the generator E.M.F. has a waveform of the type described, that is, zero up to a certain time and then of a defined waveform for all time after.

It is reasonable to suppose, if the circuit has previously had no E.M.Fs. impressed on it and is therefore quite "dead," that if the E.M.F. $e$ is applied at $t=0$, the resulting current will have a waveform which will be the same shape every time such a test is made. Also, since the circuit has been defined as linear, the resulting current waveform shape must be independent of the amplitude of the applied E.M.F. This current waveform will be given by the equation 3 for this mesh, but since our E.M.F. is defined only for $t>0$, *we must assume the solution to apply to the time* $t>0$ *only.* Thus suppose the E.M.F. has the form of Fig. 4 (b), then the current is given by :

$$L\frac{di}{dt}+Ri+\frac{1}{C}\int idt=E\cos(\omega_0 t+\phi) \quad . \quad . \quad . \quad (20a)$$

[$t>0$ only].

We can solve this equation on the assumption that the driving E.M.F. applies for all values of $t$, and not merely for $t>0$, by the method, given in Sec. 3 in this chapter, of taking some exponential form for $i$, since in this instance the E.M.F. is a simple cosine wave. This solution will give the waveform of the current $i$ as though the E.M.F. were a *continuous* cosine wave $E\cos(\omega_0 t+\phi)$. If we then merely neglect that part of $i$ lying to the left of $t=0$ we shall not obtain the correct response current to the E.M.F. applied suddenly at $t=0$, as illustrated in Fig. 4 (b), because a "discontinuity" in the

wave is involved at the instant $t=0$. Thus for finding the response waveforms in such cases, where the applied driving function is discontinuous at $t=0$, two distinct solutions of the differential equation must be made—the first solution may be obtained by assuming that the driving function is continuous for all values of $t$ and then neglecting the result for $t<0$, and the second solution must in some way derive the response waveform due to the discontinuity of the driving wave at $t=0$ and take into account all that has taken place in the mesh up to that instant.

These two solutions, added together by the Superposition principle, will then give the complete waveform of the response of the mesh to a driving function suddenly applied at the instant $t=0$. They are commonly known in mathematical texts as, respectively, the Particular Integral and the Complementary Function, and represent the *steady-state* and the *transient* behaviour of the mesh. We shall take now the second of these two cases, since we have already discussed the steady state.

## 5. The force-free response of a mesh

The E.M.F. which we have used here is defined only for the time $t>0$, but it is an *applied* E.M.F. and there is nothing assumed which excludes the possibility of there being energy stored in the circuit before $t=0$ which would not appear in the above equation 20 at all.

For instance, the E.M.F. could have been applied at the time $t=0$ by the opening of a switch across a generator, as illustrated in Fig. 5 (a). If the switch $S$ is opened at time $t=0$ the mesh is then made identical with the mesh in Fig. 1 (a),* but the current round this mesh may possibly be modified by the existence of an *initial current* $I_0$ and an *initial charge* $Q_0$, these values being known at the instant $t=0$. The energies due to $I_0$ and $Q_0$ are of two different kinds and are quite independent of one another. By this statement we mean that one gives magnetic energy in the inductance and the other electrostatic energy in the condenser, the magnitudes being respectively :

$$\tfrac{1}{2}LI_0^2 \qquad \text{and} \qquad \frac{\tfrac{1}{2}Q_0^2}{C}$$

and that these values could be specified separately.

* All the results of this section may be applied to the circuit in Fig. 1 (b), since this is the *dual* of the mesh in Fig. 1 (a), by using the rules for duals given in Sec. 2.

Now consider for a moment these two energies to be existing at $t=0$ in the mesh of Fig. 5 (b) ; if these energies be "released" at this time a current will flow round the mesh having a certain waveform, which will depend (in its shape) only on the circuit configuration, or arrangement of elements. The magnitude* of the current will clearly depend on the magnitude of $I_0$ and $Q_0$, but since the circuit is linear the waveform shape will be independent of these magnitudes. Adding up the instantaneous voltages around this mesh gives us the equation

$$L\frac{di}{dt}+Ri+\frac{1}{C}\int idt=0 \quad . \quad . \quad . \quad . \quad (20b)$$

which is the same as the mesh equation 3 except that the applied E.M.F. is now zero. It is seen that $I_0$ and $Q_0$ do not appear in this equation which determines the waveform of the current $i$ ; but we

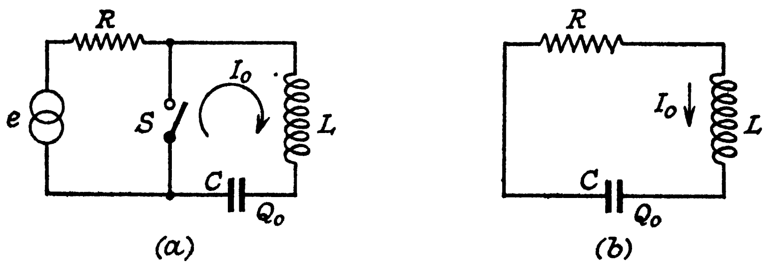



Fig. 5.—Mesh with a Switch Opened at Time $t=0$.

shall shortly see how these constants come in to determine the magnitude of the current.

Now in the mesh shown there are two sources of stored energy, but we could have considered instead a simpler mesh containing one source only, either capacitive or inductive, for example the mesh in Fig. 3 (a) (with the generator omitted). It is not possible to have more than two, since if there are a number of inductances or condensers in a single mesh we may lump the stored energies together into these two primary kinds.

Although a mesh may contain both a capacity and an inductance we may consider the case in which, at $t=0$, either one element or the other, but not both, may possess stored energy. If this energy be released at $t=0$, then a current will flow if the energy comes from

* By "magnitude" here is meant the relative amplitude of a current waveform in comparison to other currents of the same waveform.

either element, and it is reasonable to suppose that the waveform of the current will depend only on the relative magnitudes of the elements $L$, $C$, and $R$. If both storage elements possess energy at $t=0$, then each source of energy will give rise to a component current, their waveforms both depending on the values of the elements $L$, $C$, and $R$. Thus we should expect the resultant current to be divisible into two components, of similar waveforms, having relative amplitudes dependent upon the relative amounts of stored energy in the inductance and the capacity. The waveform of this resultant current is given by the "solution" of equation 20b, and interpreting this in the way discussed in the last section, we see that the "solution" may be a linear sum of a number of exponential components. Also from the preceding argument it would seem that here there are two components, the magnitudes of which both depend on $I_0$ and $Q_0$, so that $i$ has the form:

$$i=I_1\epsilon^{m_1t}+I_2\epsilon^{m_2t} \quad \ldots \ldots \quad (21)$$

where $I_1$ and $I_2$ are *arbitrary constants* which will eventually be decided by the values of $I_0$ and $Q_0$.

Since these component currents flow in the mesh freely, under no externally impressed E.M.F., the complete mesh current is sometimes called a *free oscillation*, and $i$ is called the *force-free* solution to the differential equation. The two component currents, however, are individually more important and may be said to represent the natural behaviour of the mesh, so they are known as the *natural or normal modes* (of oscillation) of the particular mesh. We have already stated that the number of degrees of freedom possessed by a complete network is equal to the number of independent meshes contained in it in which independent currents may flow. Then each of these meshes may contain either one or two kinds of energy storage element (inductive or capacitive) and so each mesh may have either one or two natural modes. Thus the total number of natural modes possessed by a complete network is determined.

Although the term "oscillation" has been used above with reference to these natural modes, they need not necessarily be oscillatory, in the sense of a periodic reversal of the current sign. To determine the actual waveform of the free oscillation for a particular mesh, the mesh equation 20b must be solved. We shall not repeat the arguments concerning the necessary exponential nature of this solution, but it can be said that, by the reasoning of Sec. 3, an

exponential form of current, $i$, enables us to write equation 20b as :

$$i\left[mL+R+\frac{1}{mC}\right]=0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

This is similar to equation 11 except that the driving E.M.F., $e$, is now zero.

One solution is obtained by writing $i=0$ but this simply implies that the circuit is quite static, or at rest, and no current is flowing.

Alternatively we have :

$$\left[mL+R+\frac{1}{mC}\right]=0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

which is a quadratic equation, and may be written :

$$(m-m_1)(m-m_2)=0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (24)$$

where

$$\left.\begin{matrix} m_1 \\ m_2 \end{matrix}\right| = -\frac{R}{2L}\pm\sqrt{\left(\frac{R^2}{4L^2}-\frac{1}{LC}\right)} \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

giving the two values of $m$, that is $m_1$ and $m_2$, which determine the waveform of the force-free current as expressed by equation 21. Thus the solution of the second order differential equation 20b gives mathematical support to our physical argument in favour of two component exponential currents or "natural modes" in this mesh containing two kinds of storage element. Substituting for $m_1$ and $m_2$ from equation 25 into 21 gives the waveforms of the two natural modes :

$$i=I_1.\epsilon^{\left[-\frac{R}{2L}+\sqrt{\left(\frac{R^2}{4L^2}-\frac{1}{LC}\right)}\right]t}+I_2.\epsilon^{\left[-\frac{R}{2L}-\sqrt{\left(\frac{R^2}{4L^2}-\frac{1}{LC}\right)}\right]t} \quad . \quad (26)$$

the total current $i$ being the force-free current in the mesh.

It is important to appreciate that this solution exists only for the time after $t=0$, since we have been determining the current which would result from the conditions $i=I_0$ in the inductance and $q=Q_0$ in the condenser at $t=0$ without knowing what the conditions were prior to this time, so that we cannot say what the mesh current waveform could be before this time.

The magnitudes of the two component current waveforms in equation 26 may now be found by inserting the given initial conditions at $t=0$. Then putting in the initial value of the current in equation 26 gives :

$$I_1+I_2=I_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

But we cannot directly put the value of the charge, $Q_0$, in this same equation (26) since the charge $q$ in the condenser is not obtainable by integrating the current given by this equation, because, as we have seen, this equation does not apply to the time prior to $t=0$. The value of $q$ at every instant including $t=0$ must, however, be implied by the original differential equation 20b and is of course contained in the term $\int i dt$, which equals $Q_0$ at $t=0$.

Taking the form of the current $i$ as given by equation 26 and inserting $i$ and $di/dt$ at the instant $t=0$ in equation 20b we have:

$$L[m_1 I_1 + m_2 I_2] + R[I_1 + I_2] + \frac{Q_0}{C} = 0 \quad . \quad . \quad . \quad (28)$$

using $m_1$ and $m_2$ again for brevity (defined by equation 25).

From these two equations (27 and 28) we may find the initial amplitudes $I_1$ and $I_2$, as :

$$\left. \begin{aligned} I_1 &= -\frac{I_0(m_2 L + R) + Q_0/C}{(m_1 - m_2)L} \\ I_2 &= +\frac{I_0(m_1 L + R) + Q_0/C}{(m_1 - m_2)L} \end{aligned} \right\} \quad . \quad . \quad . \quad . \quad (29)$$

Thus the freely oscillating current waveform is completely determined in the mesh containing one of each kind of element $L$ and $C$, resulting from given conditions of stored energy at a certain instant of time, $t=0$. We need not, in this introduction, deal with different examples of initial conditions, for instance when either $I_0$ or $Q_0$ are zero or when one element is missing, since the process of determining the current is always the same in principle ; if there are two kinds of storage element in the mesh, then the mesh equation will be of the second order, there will be two simultaneous natural modes, and hence two "arbitrary constants" will have to be found to decide their amplitudes, and these constants may always be determined by the amount of energy in the two storage elements at the time $t=0$.

Similarly if there is only one storage element in the mesh, either inductive or capacitative (for example as in Fig. 7), then there will be only one natural mode, and one arbitrary constant or amplitude to be determined from the quantity of energy stored in this single storage element.

The reader may readily prove for himself that the force-free currents in the simple meshes of Fig. 7 (a) and (b) that result from the

release of energy at $t=0$ stored in, respectively, the inductance and the capacity are given by :

$$\left.\begin{aligned}&\text{(a)}\ i=I_L.\epsilon^{-\frac{R}{L}.t}\\&\text{(b)}\ i=I_C.\epsilon^{-\frac{1}{RC}.t}\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (30)$$

The values of the amplitudes $I_L$ and $I_C$, which are the "arbitrary constants" in these cases, may be found in terms of $I_0$ or $Q_0$, the initial current or charge, in exactly the same way as for the case of the two storage-element mesh just considered.

Having worked out the analytical form of the free oscillation current and its two component natural modes (for the case of two kinds of storage element per mesh) or its single natural mode (for the case of a single storage element), let us now examine briefly the geometric forms of these currents.

We argued, in support of equation 21, that the waveforms of the two component natural modes must be similar, and on the assumption that this equation was correct we deduced these waveforms, as given by equation 26. The two indices of $\epsilon$ in this equation are composed of the same terms and are similar, apart from the change of one sign. To consider the significance of this, we must take three cases separately.

*(a) The mesh in which R is greater than* $2\sqrt{(L/C)}$

The type of current waveform which results from a sudden release of stored energy in the mesh will depend on the preponderance of the storage over the dissipation qualities of the mesh. For instance, if the capacity be charged and the charge released at $t=0$, a current will start to flow, which will transfer energy at a certain rate to the inductance ; but in so doing energy is absorbed by the resistance element. If this resistance be relatively large, a large degree of absorption of energy results and the transfer of the energy will die away at a relatively rapid rate, the current steadily falling to zero. It may be seen from equation 26 that if $R$ is greater than $2\sqrt{(L/C)}$ the two natural modes have real exponential waveforms, so that the resulting current $i$ is monotonic, that is, its direction of flow in the mesh does not reverse.

For simplicity let us write, in equation 26 :

$$\frac{R}{2L}=\delta \quad \text{and} \quad \left(\frac{R^2}{4L^2}-\frac{1}{LC}\right)=\beta^2 \quad . \quad . \quad . \quad . \quad (31)$$

We shall call $1/\delta$ the *time constant* and $\beta$ the *damped angular frequency* for reasons given later in paragraph (*c*). Then in this particular case $\beta^2$ will be positive and the expression for the current, $i$, becomes (from equation 26) :

$$\left.\begin{aligned} i &= I_1\epsilon^{-(\delta-\beta)t} + I_2\epsilon^{-(\delta+\beta)t} \\ \text{and} \qquad m_1 &= -(\delta-\beta), \quad m_2 = -(\delta+\beta) \end{aligned}\right\} \quad (32)$$

It can be seen in this case where $R > 2\sqrt{(L/C)}$ that the indices $(\delta-\beta)$ and $(\delta+\beta)$, above, must be positive, because $\delta > \beta$, from equation 31. Hence these two component natural modes will be exponential in shape, having different rates of decay. Fig. 6 illustrates their general form ; the amplitudes start at the values $I_1$ and $I_2$ at $t=0$ and decay to zero exponentially. The resultant current

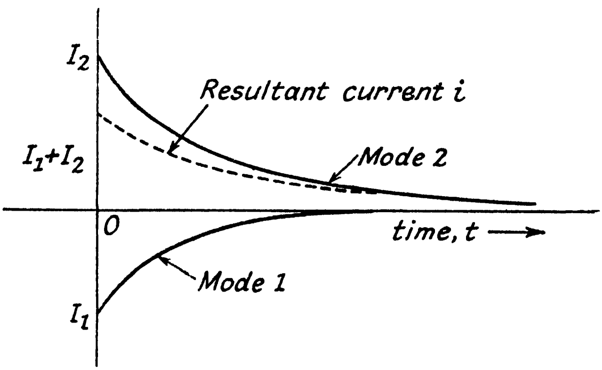



Fig. 6.—Natural Modes and Resultant "Force-free" Current in the Mesh of Fig 5 (b) for the Case $R > 2\sqrt{L/C}$.

$i$, also shown in Fig. 6 by the dotted curve, is the sum of these two natural modes, and may be written as a single term, by putting the values of $I_1$ and $I_2$ from equation 29 into 32 in terms of $\delta$ and $\beta$ :

$$i = -\left[\frac{I_0(-\delta-\beta)L + I_0R + Q_0/C}{2\beta L}\right]\epsilon^{-(\delta-\beta)t} + \left[\frac{I_0(-\delta+\beta)L + I_0R + Q_0/C}{2\beta L}\right]\epsilon^{-(\delta+\beta)t}$$

which simplifies to

$$i = \left[I_0\cosh\beta t - \left(I_0\frac{\delta}{\beta} + \frac{Q_0}{\beta LC}\right)\sinh\beta t\right]\epsilon^{-\delta t} \quad . \quad . \quad (33)$$

This expression may look more complicated than the one in equation 32, but in this form (33) it includes the initial conditions

of stored energy, in terms of $I_0$ and $Q_0$, and gives the magnitude as well as the shape of the current wave. In this equation (33) we could put either $I_0=0$ or $Q_0=0$ to determine the current wave in the case of zero energy in either the inductance or the condenser at $t=0$.

*(b) The mesh in which $R=2\sqrt{(L/C)}$*

Examination of equation 26 shows us that if $R$ is made to have the value $2\sqrt{(L/C)}$ then the square root term in the indices vanishes and the complete force-free current becomes :

$$i=I_1\epsilon^{-\frac{R}{2L}t}+I_2\epsilon^{-\frac{R}{2L}t} \quad . \quad . \quad . \quad . \quad . \quad (34)$$

That is to say, the two natural mode component currents have the same waveform shape, both being exponential with the same rate

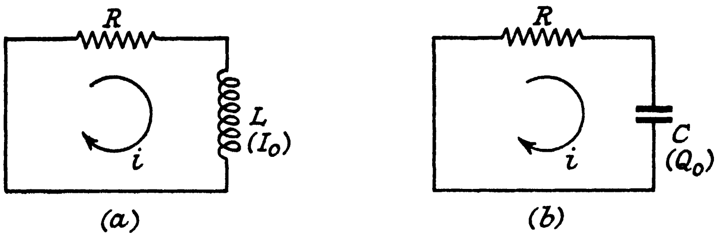



Fig. 7.—Simple Meshes with (a) Inductive Stored Energy and (b) Capacitive Stored Energy at $t=0$.

of decay. This condition in the mesh may be arrived at by considering $R$ to be a variable resistance, with its value originally set to be greater than $2\sqrt{(L/C)}$ so that the force-free current flows according to condition (*a*) above, and then its value reduced until $R=2\sqrt{(L/C)}$. Thus the term $\beta$ in the expression (32) for the current $i$ reduces to zero.

The actual waveform of the complete current $i$ may then be found from equation 33 by letting $\beta\to0$. In this case

$$\cosh\beta t\to1 \text{ and } \sinh\beta t\to\beta t$$

and then

$$i=\left[I_0-\left(I_0\delta+\frac{Q_0}{LC}\right)t\right]\epsilon^{-\delta.t} \quad . \quad . \quad . \quad . \quad (35)$$

In this limiting case the current is still not oscillatory, but flows in one direction only, the dissipation in the resistance element $R$ being such that the stored energy is all absorbed during one exchange between the capacity and the inductance.

The waveform of $i$ in this case, illustrated by Fig. 8 (a), falls to zero amplitude at the maximum possible rate consistent with its

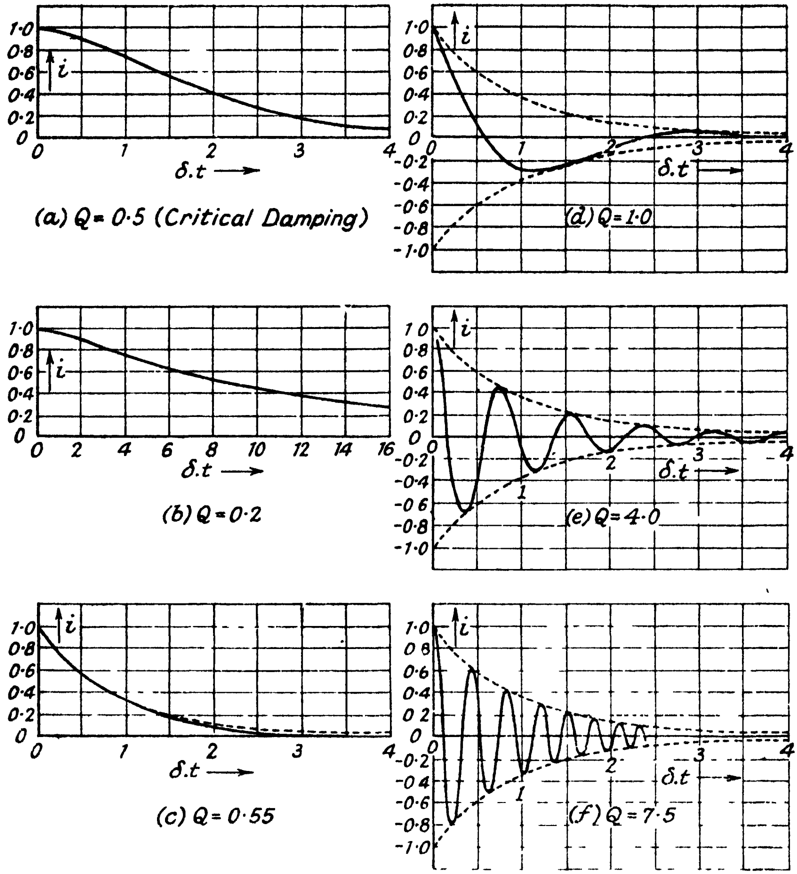



Fig. 8.—The "Force-free" Oscillations of a Series $R$, $L$, and $C$ Circuit for Different Values of Q [as in Fig. 1 (a)].

being unidirectional, and for this reason this value of $R=2\sqrt{(L/C)}$ is said to give *critical damping* of the mesh.

(c) *The mesh in which R is less than* $2\sqrt{(L/C)}$

If the value of the resistance $R$ now be reduced below this critical value, thereby reducing the dissipation as the initial stored energy in the inductance transfers to the capacity, then all the energy will not be lost during this single passage from one element to the other,

but will be transferred back to the first element, the inductance. In so doing more energy is lost in the resistance, but it may be shown that this "oscillation" of energy never ceases completely, but that the amount of energy gradually reduces with time, at a rate dependent on the value of the resistance which is absorbing the energy.

Such a brief physical argument is purely illustrative, and to show the exact form of the oscillating current we must revert to the general equation (26) for the current in the mesh of Fig. 5 (b), putting in the condition now that $R<2\sqrt{(L/C)}$.

Then the term in the index $(R^2/4L^2-1/LC)$ which we have called $\beta^2$ becomes negative, and so its square root becomes imaginary. Hence we can write equation 26 as :

$$\begin{aligned} i &= I_1\epsilon^{(-\delta+j\beta)t}+I_2\epsilon^{(-\delta-j\beta)t} \qquad \text{(where } j^2=-1) \\ &= (I_1\epsilon^{j\beta t}+I_2\epsilon^{-j\beta t})\,.\,\epsilon^{-\delta.t} \qquad \ldots\ldots\ldots\ldots \qquad (36) \end{aligned}$$

using $\delta$ as before, defined by equation 31, but taking $\beta$ to be the *magnitude* of $\sqrt{(\beta^2)}$ with a positive sign.

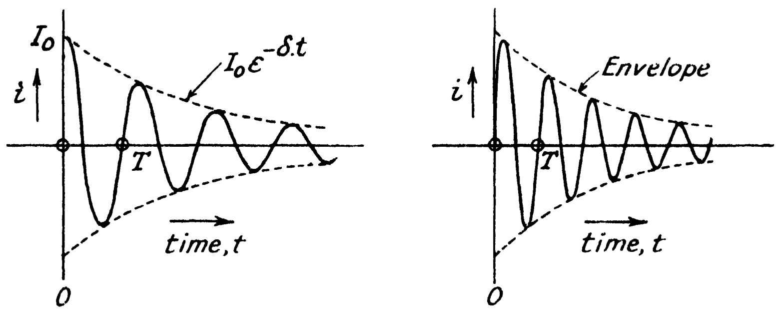
Putting in the values of the initial amplitudes $I_1$ and $I_2$ from equation 29 :

$$i=\left[I_0 \cos \beta t-\left(I_0\frac{\delta}{\beta}+\frac{Q_0}{\beta LC}\right) \sin \beta t\right]\epsilon^{-\delta t}. \quad . \quad . \quad (37)$$

which gives the waveform of the current in this case of "under-damping" of the mesh. This current is oscillatory in the true sense of the word, its direction reversing periodically while its amplitude decreases as illustrated by Fig. 9 (a) and (b). The dotted line in these figures is the *envelope* of the waves, that is the line which defines the amplitude variation of the oscillatory wave. The law of this envelope curve is given by the exponential amplitude coefficient of equation 37, that is $\epsilon^{-\delta.t}$. Thus the constant $1/\delta$ determines the rate at which the envelope reduces in amplitude with time and has already been termed the *time-constant* ; clearly the time interval $t=1/\delta$ has to elapse, from $t=0$, before this envelope falls to $1/\epsilon$ of its value at $t=0$.

The two diagrams, (a) and (b) in Fig. 9, illustrate different conditions. Firstly, there is shown the effect of different initial energy storage conditions. In (a) the condition $Q_0=0$ (at $t=0$) gives a resulting force-free current which starts at the maximum value $I_0$ ; second, in (b) the condition $I_0=0$ gives a current which starts at

zero amplitude. With different initial values of current and charge, every intermediate value of $i$ at $t=0$ is possible. There is another difference between the currents shown in (a) and (b), in that the number of complete reversals of current in a given interval of time is shown to be more in the diagram (b). The time taken for a complete reversal depends on the constant $\beta$ and may be called a *period* or *cycle time*, but this is not a constant for the *peaks* or *maxima* of a wave, such as those in this Fig. 9, in which the envelope or amplitude varies, as may be proved by differentiating ; we shall be referring to this point later in Chapter 7. The peak period will tend to become constant if $\delta$ is very small, or after a very long time, so that the variation in the wave amplitude between successive




*(a) Wave when $\beta$ is Small and $Q_0=0$*

*(b) Wave when $\beta$ is Large and $I_0=0$*

Fig. 9.—"Force-free" Current in the Mesh of Fig. 5 (b) for the Case of $R<2\sqrt{L/C}$, for Two Different Initial Conditions and Different Relations between Periodicity and Decrement.

cycles is very small. Then the period measured between peaks will become the same as the period measured between the zero's :

$$t_0=\frac{2\pi}{\beta}=\frac{2\pi}{\sqrt{(1/LC-R^2/4L^2)}} \quad . \quad . \quad . \quad (38)$$

as we can see from equation 37. The first period after $t=0$ in the waves in Fig. 9 is indicated by the point $T$. This time is measured between the *zero values* of the wave and is given by $T=t_0$ and is the same for every period.

In Fig. 9 (b) more cycles are shown for a given drop in the wave envelope amplitude than in Fig. 9 (a). The number is dependent on the relation between the time constant $1/\delta$ and the angular frequency $\beta$ and increases with the ratio $\beta/\delta$. In terms of the mesh

elements *R*, *L*, and *C* we may express this factor (which is a measure of the *continuity* of the freely oscillating current) as :

$$\beta/\delta=\frac{\sqrt{(1/LC-R^2/4L^2)}}{R/2L}$$

or

$$\beta^2/\delta^2=\left(\frac{2}{R}\sqrt{\frac{L}{C}}\right)^2-1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (39)$$

The *waveform shape* of the mesh-free oscillations then depends (see equation 37) only on this ratio $\beta/\delta$. The term $(1/R)\sqrt{(L/C)}$ is thus a measure of the damping of the mesh of Fig. 5 (b) and is commonly given the symbol Q* (not to be confused with $Q_0$ the initial charge in the condenser). The lower the value of $R$ the more oscillations will occur during the time that the envelope amplitude falls by a given amount. Clearly too this number of oscillations will increase with the ratio $L/C$. This term Q is the most compact and convenient expression for relating the dissipative to storage qualities of the mesh, and a similar expression may be used for meshes with one storage element only, such as those of Fig. 3 (a) and (b), which cannot have oscillatory force-free currents in them. In these cases the Q is normally put in terms of the response current to a driving E.M.F. of continuous sine-wave form of a given angular frequency $\omega$ such as the current of equation 16. The Q values of the meshes of Fig. 3 (a) and (b) are then, respectively, $Q=\omega L/R$ and $\omega CR$. This meaning of Q is readily lined up with our definition of $Q=(1/R)\sqrt{(L/C)}$ for a tuned circuit, if we imagine the non-oscillatory circuit of Fig. 3 (a) to be made oscillatory by a series capacity $C$ or that of Fig. 3 (b) by a shunt inductance $L$, to an angular frequency $\omega$, where $\omega=1/\sqrt{(LC)}$. Then the Q values for these circuits will be $\omega L/R$ and $\omega CR$ if the added elements are lossless.

In Fig. 8 the waveforms of the "free oscillations" of some *R*, *L*, *C* tuned circuits have been plotted, for different values of Q (circuit as in Fig. 5 (b)). The case (a) is for critical damping, $Q=0{\cdot}5$. For cases of $Q>0{\cdot}5$ the waveforms are oscillatory, and the exponential envelopes have been shown dotted. The initial conditions ($t=0$) chosen here are $i=I_0$ and $Q_0=0$. For a mesh of low Q value the envelope decay rate more nearly approaches the slope of the sinusoidal wave itself, as in Fig. 8 (c) and (d), so that there

* For the shunt tuned circuit, Fig. 1 (b), Q may be defined as the inverse of this, or $1/Q=(1/R)\sqrt{(L/C)}$, as follows from solution of equation 4 for this circuit.

are fewer oscillations during a given drop in the envelope amplitude. For a mesh of even lower Q the envelope decay rate may, so to speak, overtake the slope of the oscillating wave itself, which then becomes non-oscillatory and of the form of Fig. 8 (a) and (b). The critical value of Q, corresponding to critical damping of the mesh, where $R=2\sqrt{(L/C)}$ is clearly $Q=(1/R)\sqrt{(L/C)}=0{\cdot}5$.

If the resistance element $R$ were to vanish altogether, then the mesh Q would become infinite and the force-free oscillations would continue for ever without decrement, as illustrated in Fig. 10. If $R=0$ then the time-constant $1/\delta$ is infinity, while $\beta=1/\sqrt{(LC)}$, and

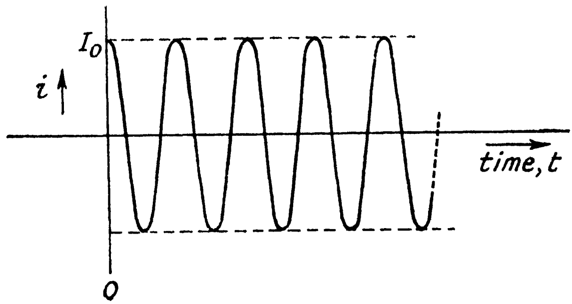



Fig. 10.—"Force-free" Current in the Mesh of Fig. 5 (b), when $R=0$.

then the expression for this sinusoidal wave of constant amplitude is given, from equation (37), by:

$$i=\left[I_0\cos\frac{t}{\sqrt{(LC)}}-\frac{Q_0}{\sqrt{(LC)}}\sin\frac{t}{\sqrt{(LC)}}\right] \quad . \quad . \quad . \quad (40)$$

Clearly the initial phase of this wave, at time $t=0$, depends on the relative values of $Q_0/\sqrt{(LC)}$ and $I_0$ and in Figure 10, the case has been illustrated in which $Q_0=0$, so that the oscillation starts off with its maximum value $I_0$.

In this case of zero dissipation the wave envelope is a continuous straight line, and hence the periodic time of the sinusoidal current wave is constant and equal to $2\pi\sqrt{(LC)}$ measured at the peaks or zeros of the wave. The *frequency* of this wave, or the number of cycles per second is hence the inverse of this,

$$\text{Steady frequency } f=\frac{1}{2\pi\sqrt{(LC)}} \quad . \quad . \quad . \quad . \quad (41)$$

## 6. Modulated waves and envelope waves

When dealing with the subject of the transient response of networks and the geometry of waveforms, we shall continually find the need to refer to two basic types of wave. First, the modulated wave, for instance that expressed by equation 37, which is a periodic oscillation having the exponential envelope form $\epsilon^{-\delta.t}$, being the force-free response of the $R$, $L$, $C$ mesh of Fig. 5 (b). Second, the envelope wave of equation 30a, of the form $\epsilon^{-2\delta.t}$, which is the force-free oscillation of the $R$, $L$ mesh of Fig. 3 (a).

We have shown in the preceding section how these two types of wave are "characteristic" of these two types of mesh, one containing an inductance and a capacity, while the other contains an inductance only, as a "storage element." It is true that the oscillatory character of the response of the $R$, $L$, $C$ mesh of Fig. 5 (b) is destroyed if the mesh Q is less than the critical value 0·5, but let us confine our attention, at the moment, to cases of higher Q. Then comparison of these two meshes and their force-free responses shows that the effect of removing the capacity from the $R$, $L$, $C$ mesh changes its response from the modulated wave with the envelope form $\epsilon^{-\delta.t}$ to a simple envelope waveform $\epsilon^{-2\delta t}$—that is, the envelope of the $R$, $L$, $C$ mesh waveform but with half the time-constant. In order to produce an $R$, $L$ mesh with a force-free response identical with the $R$, $L$, $C$ mesh response envelope, the value of $L$ must be *doubled.*

Exactly the same principle may be applied to the shunt $R$, $L$, $C$ circuit between the terminals $AB$ in Fig. 1 (b), which may readily be shown to have a force-free voltage response of a modulated sinusoidal waveform (again similar to that in Fig. 9) and the $R$, $C$ circuit between terminals $AB$ in Fig. 3 (b). Then the response of the $R$, $C$ circuit is an envelope wave identical with the envelope of the modulated wave response of the $R$, $L$, $C$ circuit, provided that, in this case, the value of $C$ *be halved.*

This is an important theorem, the usefulness of which we shall discuss in a later chapter,* and it is mentioned here in order to emphasise this relation between the elementary meshes containing $R$, $L$, and $C$ elements and those containing $R$ and $L$ or $R$ and $C$ only. If such meshes are built up into more complex networks the same general idea may be applied, and networks which can be split up entirely into $R$, $L$, $C$ branches will have oscillatory force-free responses consisting of sinusoidal waves modulated by some

* See Chapter 3, Sec. 33, "low-pass/band-pass analogy."

envelope form, not necessarily exponential, while networks composed entirely of $R$, $L$ or $R$, $C$ branches will have non-oscillatory responses. Of course there are numbers of networks in common use which are not strictly composed purely of either type of branch and their responses may consist of a mixture of modulated waves and envelope waves, but the broad division that has been indicated here is an important one, and we shall be referring frequently in this book to these two types of wave—the modulated sinusoidal wave, with its own envelope form, and the simple envelope wave.

## 7. The forced response and the complete response of a mesh

In the last two sections we have been dealing with the first of the two components of current which flow in a mesh as a result of applying suddenly a driving E.M.F. at the instant $t=0$. This current has been shown to be a "force-free" current having a waveform characteristic of the mesh configuration, but dependent only, in amplitude and initial value, on the energy stored in the mesh at the instant $t=0$; its form was deduced by solving the equation 20b (for the general $R$, $L$, $C$ mesh) which does not involve the expression for the driving E.M.F.

In order to find the form of the total resulting current, due to the sudden application of a driving E.M.F. to the network terminals, we must add a second current component to this first force-free solution, and this second current is given by the solution of the equation 20a but without restriction as to time—that is, it will be the current which would result from the application of the *continuous* driving E.M.F., $E \cos(\omega_0 t+\phi)$. Having determined the current for this particular form of E.M.F. we then ignore its value for $t<0$. This process is the same whatever the form of this driving E.M.F., and the resulting current may very occasionally be found by inspection, but only too often this is not possible and methods have had to be evolved for solving differential equations of this type. It would be out of place to discuss these here, since we are not proposing to deal with operational methods ; the reader is probably familiar with the use of the operator[6] D, and may be acquainted with the methods of Heaviside[4, 10, 11] and Carson[12, 13] and others,[14, 15, 16] which all are methods of solving the differential equations of a network. Some operational methods solve the equation with the inclusion of the initial conditions, so that two solutions such as we have discussed here are not required. Unfortunately these methods are in general mathematically difficult

and are more the affair of the mathematician than the engineer, who has to concern himself with so many other aspects of the subject. For the benefit of those readers particularly interested, the more practical books on the subject are indicated in the list of references.

It is thought sufficient to have indicated here the general method of dealing with a discontinuous driving E.M.F., by solving for the steady-state condition as though the particular E.M.F. had been flowing for ever, and neglecting the solution for times before $t=0$ and then taking up the effect of the discontinuity by finding the current which results from the known storage of energy at the instant $t=0$.

One important note: when the conditions at $t=0$ are inserted in the expression for the current, in order to determine the arbitrary constants of the force-free solution, these conditions must be inserted in the expression for the *complete current* (i.e. steady-state+force-free current). It is clear this must be so, since at the instant $t=0$ these stored energies are trying to dissipate themselves in the presence of the initial value of the driving E.M.F.

We have seen that the force-free component of the current dies away at a rate dependent on the time-constant of the mesh and may be considered negligibly small after a certain time. The same is true of more complicated networks having a number of meshes, so that after a length of time the transient may be considered to have vanished and the whole network behaves as though it were in a steady state and as though the driving E.M.F. had been acting for an indefinitely long time previously. All E.M.Fs. have to be switched on or applied to a network's terminals at some definite instant, and a length of time must elapse, long compared with the networks natural time-constant, before the initial transient has died away and the behaviour may be considered steady state. Such a condition is known as *quasi-steady state*, and is the practical condition, though when the terms "continuous sine-wave," "steady state," etc., are used, they are really mathematical conveniences, and may be approached in practice by considering only times long after the switching instant $t=0$.

## 8. The steady-state approach to transient problems—the idea of an impedance function

In Sec. 3 we showed how the response of a mesh to an applied E.M.F. (of given waveform) could be calculated by splitting up this E.M.F. waveform into a sum of elementary waves of exponential

form. The actual analytical methods of deriving these elementary exponentials from a wave of given form have not yet been indicated, but are the concern of the next chapter ; so far it has merely been shown that such a series of exponentials satisfies the types of differential equations that arise in network analysis. The important distinction that must be made between this method of solution and the other method discussed, by which we may calculate the response to a wave whose form is defined only for the time after an instant $t=0$, is that the former method is directly applicable to *continuous* waves only. We can use this former method only if the waveform of the applied E.M.F. is defined for all time from $t=-\infty$ to $+\infty$, so that the question of stored energy does not arise.

But waves of any shape may be regarded, in this way, as being composed of a number of exponential component waves, even including waves which are of zero amplitude up to time $t=0$, such as are illustrated in Fig. 4. These waves appear to be discontinuous at the instant $t=0$, but their shape may be approached as closely as desired by taking a sufficient number of exponential components, and indeed it is possible to deal with an infinite number, as we shall see later when we come to use Fourier integrals and deal with the waveforms of transients in more detail.

Now sometimes the discontinuity in an applied wave is produced by a switching action or by some means which suddenly changes the circuit configuration. In these cases it is usual to calculate the response currents by considering the applied wave after the instant $t=0$ and the transient due to the discontinuity, but it is possible to simulate this switching action by applying an appropriate E.M.F., so that even these problems involving changes of configuration may, if desired, be reduced to steady-state problems which may be tackled by dividing up the applied waveform into its complete equivalent series of exponential waves. It is not to be implied that such a method is necessarily the best, but these two general methods of approach should be appreciated.

In conclusion let us consider briefly another aspect of the steady-state approach to the problem of the transient response of a network. We shall again take the mesh shown in Fig. 1 (a) as an example. In Sec. 3 it was argued that the response current must be capable of being expressed as a series of exponential currents, as given by equation 13, so that the applied E.M.F., $e$, itself may be represented by a series of elementary exponential voltages, as in equation 14. These components of current and E.M.F. correspond in

waveform, term by term, as may be seen from these two equations, and each component of E.M.F. may be considered to give rise to its particular component of the response current.

The ratio of the applied E.M.F. to the response current will be (from equations 13 and 14):

$$\frac{e}{i}=\frac{I_1[m_1L+R+1/m_1C]\epsilon^{m_1t}+I_2[m_2L+R+1/m_2C]\epsilon^{m_2t}+\ .\ .\ .\ ,\text{ etc.}}{I_1\epsilon^{m_1t}+I_2\epsilon^{m_2t}+\ .\ .\ .\ ,\text{ etc.}} \quad . \quad . \quad . \quad (42)$$

which is clearly a function of time, in this general case. This ratio is known as the impedance of the mesh to *the particular current waveform i*. This generalised idea of an impedance is not of much interest, but certain special cases are in common use, for instance the impedance to a sinusoidal wave, with which we shall deal in due course. The impedance of a network to a step-wave (as in Fig. 4 (a)) is also used, particularly in certain operational methods of analysis. For an example of its use let us suppose such a wave of E.M.F. to be impressed on the mesh of Fig. 1 (a) and the response current waveform to be known (by oscillographic measurement or otherwise—the method does not concern us here). Then the instantaneous ratio of these waveforms $e/i$ is a specification of the response of this particular mesh to a step-wave E.M.F. of *any arbitrary amplitude*, provided that the mesh is linear. This ratio is called the *indicial impedance* or its inverse the *indicial admittance* of the mesh[12], provided it refers to an applied step-waveform. Knowing this indicial impedance, which of course is a function of time, for our particular mesh, we may calculate the response of this mesh to a wave of any other form by means of the Superposition Theorem, as will be explained in Chapter 4.

The response of the mesh to an applied E.M.F. with a purely exponential waveform gives us the simplest kind of impedance function. As we saw from equations 10 and 11, the assumption of a single exponential response term in the differential equation of the mesh implies that the driving E.M.F. itself must be purely exponential. The ratio of E.M.F. to response current then becomes:

$$\frac{e}{i}=\frac{I[mL+R+1/mC]\epsilon^{mt}}{I\epsilon^{mt}}=[mL+R+1/mC] \quad . \quad (43)$$

to which expression we referred, in this previous instance, as the "network-function" (see equation 9). In this instance it is the impedance of the mesh *to an exponential current* and is constant,

in contrast to the impedance to a general current wave (equation 42), which is a time function. Hence, to specify the response of this particular mesh to an exponential wave, a certain constant is required. Again, knowing the value of this constant for every exponential component in a more complicated wave, which may be represented by such a series as that in equation 14, the response to this wave may be calculated by simple addition according to the Superposition Theorem.

The most important and commonly used impedance form is derived from a special case of the exponential wave, as was suggested in Sec. 3, by taking imaginary values for the constant index $m$. In equation 15 the current was written as $i=I.\epsilon^{j\omega t}$, which represents a continuous sinusoidal waveform. This imaginary index is quite admissable, since this expression for the current satisfies the differential equation, but a more direct approach to the idea of a sinusoidal response may be made by direct substitution of this form for the current:

$$i=I\cos\omega t \qquad \text{(as in equation 16)}$$

in the differential equation (3) for the mesh.

This gives:

$$I\left[R\cos\omega t-\left(\omega L-\frac{1}{\omega C}\right)\sin\omega t\right]=e \quad . \quad . \quad (44)$$

which may be written $I'\cos(\omega t+\omega t_1)=e$ where:

$$I'=I\sqrt{[R^2+(\omega L-1/\omega C)^2]} \quad \text{and} \quad \tan\omega t_1=\left(\frac{\omega L-1/\omega C}{R}\right) . \quad (45)$$

Thus the impedance of this mesh, in this case of a continuous sine wave, is:

$$\frac{e}{i}=\frac{I'\cos(\omega t+\omega t_1)}{I\cos\omega t} \quad . \quad . \quad . \quad . \quad . \quad . \quad (46)$$

Here both the E.M.F., $e$, and current, $i$, have sinusoidal waveforms of equal period, but with a relative shift in time of $t_1$. It is true that the instantaneous ratio of $e/i$ is in general a function of time, but taking this relative shift into account the ratio becomes a constant. Then, in this instance of the response to a sinusoidal wave, *two* quantities must be specified for the impedance function:

(1) The magnitude of $e/i$, usually written $|e/i|$ and called the *modulus*, which in this example is:

$$|e/i|=\sqrt{\left[R^2+\left(\omega L-\frac{1}{\omega C}\right)^2\right]}$$

(2) The relative time shift $t_1$, which is usually specified by the angle $\omega t_1$, called the *phase angle.*

This impedance function is of course the conventional steady-state impedance to an alternating current which is used in all so-called "steady-state response" calculations, and is the basis of the vectorial method of circuit analysis. More complicated waves may, again, be represented as a series of elementary waves of this type, as in the equation 13, where $m_1$, $m_2$ . . . , etc., have imaginary values $j\omega_1, j\omega_2$ . . . , etc. Then the response of a network to such complicated waves may be calculated in terms of the response to each of these elementary sinusoidal waves, since the individual responses may be added together by the Superposition Theorem. This method is known as the Fourier analysis method, and this also will form a considerable part of the matter dealt with in subsequent chapters.

## REFERENCES

1. O. Heaviside, "Electromagnetic Theory," Vol. II, 1899, and his "Electrical Papers," 1893.
2. Steinmetz, "Complex Quantities," "Proc. Int. Elec. Congress," Chicago, Aug. 1893.
3. H. S. Carslaw, "Introduction to the Theory of Fourier Series and Integrals," Macmillan & Co., 1930
4. L. Cohen, "Heaviside's Electrical Circuit Theory," McGraw-Hill, New York, 1928.
5. Pupin, *Trans. American Inst. Elec. Eng.*, Vol. XVI, 1899, and Vol. XVII, 1900.
6. H. T. H. Piaggio, "Differential Equations," Bell & Sons, London.
7. E. A. Guillemin, "Communication Networks," Vol. I, John Wiley & Sons, New York (London, Chapman & Hall), 1931.
8. A. Bloch, "A New Approach to the Dynamics of Systems with Gyroscopic Coupling Terms," *Phil. Mag.*, Sec. 7, Vol. 35, May 1944, No. 244.
9. W. P. Mason, "Electromechanical Transducers and Wave-Filters," Van Nostrand., New York, 1942.
10. G. W. Carter, "The Simple Calculation of Electrical Transients," C.U.P., 1944.
11. E. Stephens, "Elementary Theory of Operational Mathematics," McGraw-Hill, New York, 1937
12. J. R. Carson, "Electric Circuit Theory and the Operational Calculus," McGraw-Hill, New York, 1926.
13. H. S. Carslaw and J. C. Jaeger, "Operational Methods in Applied Mathematics," O.U.P., 1941.
14. N. W. McLachlan, "Complex Variable and Operational Calculus with Technical Applications," C.U.P., 1939.
15. V. Bush, "Operational Circuit Analysis," John Wiley, New York, 1929.

16. M. F. GARDNER and J. L. BARNES, "Transients in Linear Systems (Studied by the Laplace Transformation)," John Wiley, New York, 1942 (London, Chapman & Hall).
17. W. P. MASON, "Electrical and Mechanical Analogies," *B.S.T.J.*, Vol. 20, p. 405, Oct. 1941.
18. H. F. OLSON, "Dynamical Analogies," Van Nostrand, New York (London, Chapman & Hall), 1943.

The following references also are interesting in connection with the historical and fundamental aspects of network behaviour:

(A) J. BLANCHARD, "A History of Electrical Resonance" (with included references), *B.S.T.J.*, Vol. 20, Oct. 1941.
(B) C. P. STEINMETZ, "Elementary Lectures on Electric Discharges, Waves, and Impulses," McGraw-Hill, New York, 1914.
(C) G. A. CAMPBELL, "Physical Theory of the Electric Wave Filter," *B.S.T.J.*, Vol. 1, Nov. 1922, p. 1.
(D) R. V. L. HARTLEY, "Transmission of Information," *B.S.T.J.*, Vol. 7, Aug. 1928, p. 535.
(E) S. A. SCHELKUNOFF, "The Impedance Concept and its Application to the Problems of Reflection, Refraction, Shielding, and Power Absorption," *B.S.T.J.*, Vol. 17, Jan. 1938, p. 17.
(F) B. VAN DER POL, "A New Transformation in Alternating Current Theory with an Application to the Theory of Audition," *Phil. Mag.*, Vol. VII, 1929, p. 477.
(G) N. MARCHAND, "The Response of Electrical Networks to Non-Sinusoidal Periodic Waves," *Proc.I.R.E.*, Vol. 29, No. 6, June 1941, p. 330.
(H) J. GREIG, "Pulses and Transients" (lecture before the Wireless Personnel Committee), "University Radio Conference Proceedings," published by Privy Council Office, London.

## CHAPTER 2

# THE FREQUENCY SPECTRA OF MODULATED WAVES, PULSES AND TRANSIENTS

### 9. The continuous sine wave—its common use in analysis and in practice.

Although the continuous sine wave is the most commonly used type of waveform in network analysis, it is not a "natural" shape of wave, since all oscillations have to start at some definite instant of time. Also all physical passive networks (those containing no internal energy generators) must contain some degree of dissipation, and so oscillations started in them must decay in amplitude in the way discussed in Sec. 5. This damped sine wave is a far more "natural" type of wave therefore, but is not so popular as a basis of circuit analysis, although a very powerful method rests on the response to waves of this type. The popularity of the continuous sine waveform arises from a number of obvious causes, for instance:

(1) It provides a very simple solution to the type of differential equation that arises in circuit theory, and is the basis of the elementary alternating current theory.

(2) It is a waveform which is commonly the ideal of A.C. generators; the wave is then considered to be produced by the rotation of a coil in a uniform magnetic field.

(3) The behaviour of wave filters and an immense variety of standard types of network is specified from the aspect of sine-wave response (that is, the *frequency* characteristics of networks in common use have been published widely, though the information on their response to other wave shapes is far more scanty).

For these and other reasons the interpretation of the response of networks to pulses, transients, and other complex waves in terms of their response to the simple continuous sine wave is most commonly taught and understood.

Before proceeding to the analysis of complex waves in terms of sine waves, let us consider a few aspects of the simple A.C. vector, since this idea is a very important one in connection with the geometrical approach to the subject.

## 10. The A.C. vector—trigonometrical and exponential notations

Confusion is sometimes caused by the two notations in common use in A.C. vector theory, but they are both representations of the same geometrical idea. In elementary treatments of the subject, the A.C. vector is said to be a rotating line, the projection of which on to some fixed axis then varies sinusoidally with time, and so may be regarded as "generating" the sine wave. It is imagined as lying in the plane of a coil rotating in a uniform magnetic field, so that the projection of the vector gives, at every instant, the amplitude of the E.M.F. induced in the coil.[1, 4] We need not pursue this idea here, but assuming the reader to be familiar with the elementary use of A.C. vectors, we shall proceed with explaining the relation between the two notations.

In Sec. 3 of Chapter 1 we showed that the continuous sine wave could be written in two ways:

$$(a) \quad i=I\,.\,\epsilon^{j\omega t}$$

or

$$(b) \quad i=I\cos\omega t \qquad \text{. . . . . . . . (47)}$$

and that either form was satisfactory as regards the solution of the differential equation. The form (*b*) is the trigonometrical form, and represents directly the sinusoidal waveform of the current. It is equal *to the real part* of form (*a*), since:

$$(c) \quad I\,.\,\epsilon^{j\omega t}=I\cos\omega t+jI\sin\omega t \qquad \text{. . . (47)}$$

Now consider the vector shown in Fig. 11, plotted on the conventional complex plane. This is a single vector of length $I$ rotating at an angular speed $\omega$ radians/sec. For the moment let us say that this vector lies along the real axis $OR$ at time $t=0$. Its projections at every instant on to the two perpendicular axes, are:

$$OA=I\cos\omega t$$
$$OB=I\sin\omega t \qquad \text{. . . . . . . . (48)}$$

so that this vector, being the vectorial sum of these projections, represents the quantity $I\epsilon^{j\omega t}$, since

$$OE=(OA+jOB)=I(\cos\omega t+j\sin\omega t)$$

The projection $OA$ on the real axis varies sinusoidally with time, and has been plotted in Fig. 11 as a function of time. Now with the vector in the position shown on the diagram at the instant $t=0$, the sine wave has its origin as shown, the peak being shifted by a phase angle $\phi$, which corresponds to a time of $t=\phi/\omega$ secs.

Similarly, on the vector diagram, the vector shown has its angular origin shifted forward by $\phi$ radians from the origin which we first considered $OR$.

Then with this origin as shown the rotating vector may be written as the vector sum of the projections on the two perpendicular axes:

$$i=I\{\cos(\omega t-\phi)+j\sin(\omega t-\phi)\} \quad . \; . \; . \quad (49)$$

which also is
$$i=I\epsilon^{j(\omega t-\phi)} \quad . \; . \; . \; . \; . \; . \; . \quad (50)$$

or
$$i=[I\epsilon^{-j\phi}]\,\epsilon^{j\omega t} \quad . \; . \; . \; . \; . \; . \; . \quad (51)$$

This shows us the mathematical simplicity of using the exponential notation for vectors. From equation 51 we see that the quantity $\epsilon^{j\omega t}$ has the complex amplitude $I\epsilon^{-j\phi}$ and therefore we may plot

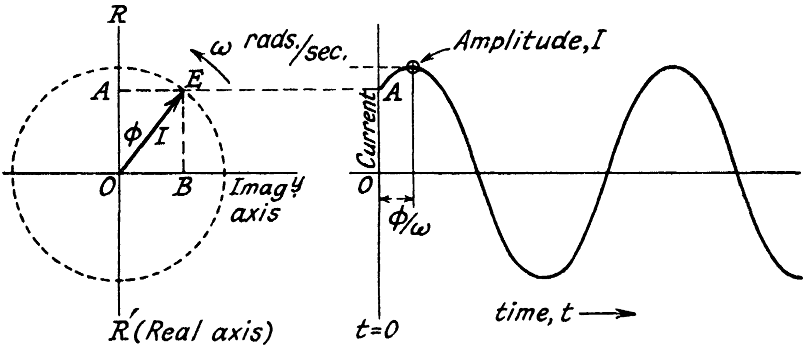



Fig. 11.—The Simple Rotating A.C. Vector.

this on the diagram, as a *stationary* vector, of length $I$ and phase angle $-\phi$ relative to the positive real axis (the chosen angular origin). This is of course the same vector shown in Fig. 11 if we omit the arrow showing it to rotate at $\omega$ radians/sec. We are, so to speak, omitting the variable $\epsilon^{j\omega t}$ in equation 51, and are understanding that the waveform represented by this vector is a pure sine wave and that its frequency is known. Vectors of similar frequency may be added together on such a diagram, since they will all be "stationary."

The comparative clumsiness of the trigonometrical notation for *analytical work* may be seen by comparing the above vector expression with the form:

$$i=I\cos(\omega t-\phi) \quad . \; . \; . \; . \; . \; . \quad (52)$$

which also represents the vector in Fig. 11 with the origin as shown there. Here the amplitude and phase are separated, and the whole

expression must be used when dealing with the vector, in analysis. We may also write this in terms of its orthogonal components:

$$i=[I\cos\phi]\cos\omega t+[I\sin\phi]\sin\omega t \quad . \quad . \quad . \quad (53)$$

Or, again, it may be represented by the terms:

$$\text{Vector } i=\sqrt{(OA^2+OB^2)} \text{ at an angle } -\tan^{-1}\frac{OB}{OA} \quad . \quad (54)$$

Now equation 53 is important in that it shows that the rotating vector may be represented by two orthogonal quantities which vary in amplitude with time. These two quantities have maximum amplitudes:

$$I\cos\phi \quad \text{and} \quad I\sin\phi \quad \text{(respectively)} \quad . \quad . \quad (55)$$

We shall be using an extension of this idea later when dealing with vectors which rotate at a non-uniform speed and also vary in length. At present these two orthogonal components give the simple uniformly rotating vector. For convenience they will be referred to as the "in-phase" and "quadrature" components of the vector.

### 11. Conjugate vectors

There is a very simple alternative way of regarding A.C. vectors, which we shall see later is also the more logical. The difficulty that some people find in reconciling the use of a complex quantity to represent a sine wave (a real quantity) may be overcome in the following way. Consider again the vector, as shown in Fig. 11, and its expression as:

$$i=I\{\cos(\omega t-\phi)+j\sin(\omega t-\phi)\} \quad . \quad . \quad . \quad (56)$$

This may be converted into the true expression for the sine wave current $i=I\cos(\omega t-\phi)$ by the addition of the *conjugate vector*:

$$i=I\{\cos(\omega t-\phi)-j\sin(\omega t-\phi)\} \quad . \quad . \quad . \quad (57)$$

and to keep the total amplitude correct *both vectors must be halved in length.* This is obvious from simple addition of equations 56 and 57. Such a conjugate pair of vectors is shown in Fig. 12; these vectors are assumed to rotate in opposite directions at angular speed $\omega$, and their projection sum on the real axis varies sinusoidally, while their projection sum on the imaginary axis is always zero. Since these vectors rotate in opposite directions, their angular speeds must be $\pm\omega$ and this gives rise to the idea of a *negative frequency.* What this means physically is not important, since the whole use of vectors is purely operational, but the use of such negative values

of frequency in calculations is quite legitimate and gives practical results if used with consistency, just as does also the use of $j$. Thus we must remember that:

$$\cos(-\omega t) = \cos \omega t \quad \text{and} \quad \sin(-\omega t) = -\sin \omega t \quad . \tag{58}$$

and there need be no confusion.

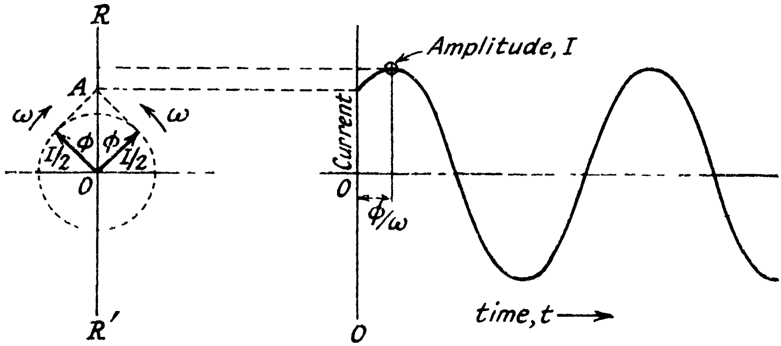



Fig. 12.—A Pair of Conjugate Vectors.

Such vectors must always be dealt with *in pairs*, each pair representing a single sine wave. Now these conjugate vectors may each be divided into two components, just as in equation 53 we have

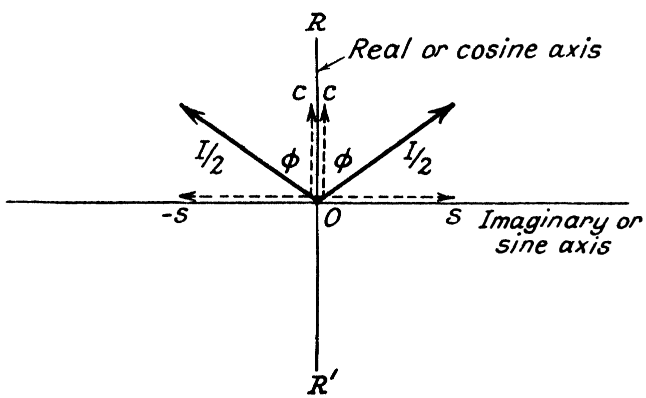



Fig. 13.—Cosine and Sine Components of Two Conjugate Vectors.

shown a single vector to be divided into what we called the "in-phase" and "quadrature" components, one a cosine wave (lying along the real axis at $t=0$) and the other a sine wave (lying along imaginary axis at $t=0$). In Fig. 13 we have shown the orthogonal

components of the conjugate vectors of Fig. 12 (and equations 56 and 57). The components lying along the real axis will be cosine waves and those on the other axis sine waves, so from this point of view we may call these axes the cosine and the sine axes respectively.

The lengths of these pairs of components will be:

$$\left.\begin{aligned} OC &= \frac{I}{2}\cos\phi \\ OS &= \frac{I}{2}\sin\phi \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (59)$$

just as in equation 55 for the single vector, though in this case the cosine component will consist of two positive vectors and the sine component will consist of one positive and one negative vector. The components of two conjugate vectors which are used to represent a sinusoidal wave in any arbitrary phase $\phi$ at ($t=0$) always go in such pairs; one pair of like sign is called an *even* or a *symmetrical* pair and one of opposite sign called an *odd* or *skew-symmetrical* pair. The even pair alone represents a cosine wave (since their sum has its maximum amplitude at $t=0$), while the odd pair represents a sine wave (their sum being zero at $t=0$).

## 12. The frequency spectrum of a simple sinusoidal wave

The simple wave shown in Fig. 11 may be represented by a "frequency characteristic" or spectrum, as a vertical line (of length

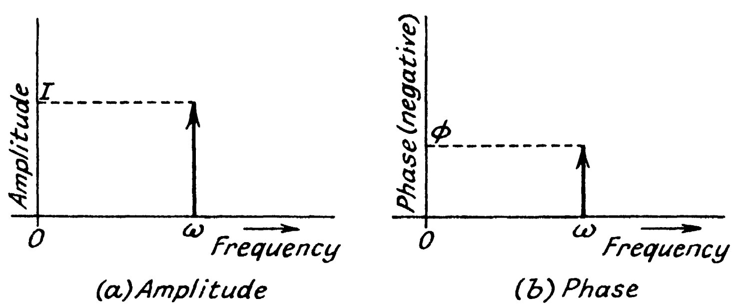



Fig. 14.—The Amplitude and Phase Spectrum of the Sinusoidal Wave $i=I\cos(\omega t-\phi)$

proportional to the wave amplitude $I$) on a frequency axis. In the most usual form, such a characteristic appears as shown in Fig. 14. To make the representation of the wave complete, two diagrams are needed, (a) showing the amplitude $I$, (b) showing the phase $\phi$.

Further it needs to be understood whether the wave represented thus is a cosine or a sine wave, i.e. the same diagram represents both

$$i=I\cos(\omega t-\phi) \quad \text{and} \quad i=I\sin(\omega t-\phi)$$

just as in the case of the vector diagram, the real axis was defined as the axis of projection (Sec. 10)—a purely arbitrary, though necessary, definition.

Alternatively a simpler and more logical way of drawing such a frequency characteristic or spectrum is by the use of the conjugate components, explained in the preceding section. The symmetrical and the skew-symmetrical pairs as we called them, being the cosine and sine components of the conjugate vectors of Fig. 13, may be plotted on a frequency axis as shown in Fig. 15. The lengths of these components will be given by equation 59—the lengths $OC$ and $OS$ in Fig. 13.

It is seen that the cosine components are plotted in Fig. 15 as a symmetrical pair of spectrum terms, that is to say as a pair of terms at the frequencies $\pm\omega$ and of equal amplitudes $OC$ of the same sign. The sine components are plotted as a skew-symmetrical pair of spectrum terms, at the frequencies $\pm\omega$ and of equal amplitudes $OS$ but of opposite signs. That these four terms together represent the current wave $i$ may be readily checked by addition. Writing **C** and **S** for the lengths $OC$ and $OS$ of the vectors, the sum of the four terms is:

$$\left.\begin{aligned}\text{Cosine component} &= [\mathbf{C}]\cos\omega t+[\mathbf{C}]\cos(-\omega t)\\ \text{Sine component} &= [\mathbf{S}]\sin\omega t+[-\mathbf{S}]\sin(-\omega t)\end{aligned}\right\} \quad . \quad . \quad (60)$$

But if $\mathbf{C}=(I/2)\cos\phi$ and $\mathbf{S}=(I/2)\sin\phi$, on a scale of length, then the total sum is:

$$\left(\frac{I}{2}\cos\phi\right)\cos\omega t+\left(\frac{I}{2}\cos\phi\right)\cos\omega t+\left(\frac{I}{2}\sin\phi\right)\sin\omega t+\left(\frac{I}{2}\sin\phi\right)\sin\omega t=I\cos(\omega t-\phi)$$

These two parts of a complete spectrum diagram (the symmetrical and skew-symmetrical parts) may very neatly be combined into a single diagram[28] by direct arithmetic addition. Thus Fig. 15 (c) shows the lengths of the positive and negative frequency components added. The component of frequency $+\omega$ has the length $(\mathbf{C}+\mathbf{S})=I/2(\cos\phi+\sin\phi)$ while the component of frequency $-\omega$ has the length $(\mathbf{C}-\mathbf{S})=I/2(\cos\phi-\sin\phi)$. Each term is an algebraic sum of a sine and a cosine wave, but if the diagram be plotted in

this compact way, the two parts may readily be separated by addition and subtraction of the lengths (C+S) and (C−S).

This method of representing a wave spectrum and vector diagram by conjugate components is really more than "just another notation." It bridges the gap between the complex exponential notation, so useful in analysis, and the trigonometrical notation, which is the more direct way of representing what is essentially a real quantity—the sinusoidal waveform. It will be shown in the next sections that the spectra of amplitude-modulated waves, as normally plotted, are in an identical form, and when we come to deal

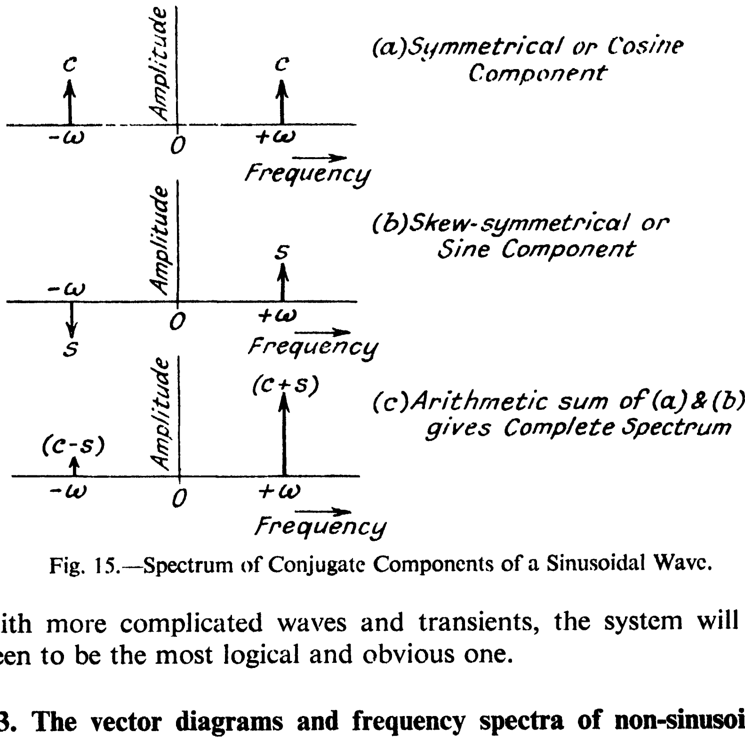



Fig. 15.—Spectrum of Conjugate Components of a Sinusoidal Wave.

with more complicated waves and transients, the system will be seen to be the most logical and obvious one.

## 13. The vector diagrams and frequency spectra of non-sinusoidal waves

Suppose we have a number of sinusoidal wave generators, such as is shown in Fig. 16, all of zero impedance and each producing its own current component in the resistance load $R$. If the E.M.Fs. and frequencies ($E_1$, $E_2$ . . . . etc., and $\omega_1$, $\omega_2$ . . . . etc.) are different, how may we express the current $i$ in the resistance as a vector diagram or as a spectrum? This current will be of a

non-sinusoidal waveform, being the sum of a number of sinusoidal waves of different frequencies (and amplitudes). In this case there is no meaning to the idea of relative phase between the waves unless a time origin be fixed. Thus, if we express the resultant current as:

$$i=I_1 \cos(\omega_1 t-\phi_1)+I_2 \cos(\omega_2 t-\phi_2)+I_3 \cos(\omega_3 t-\phi_3) \ldots \text{etc.} \quad (61)$$

where $I_n=E_n/R$, we are assuming a time origin, $t=0$, such that at this point the currents have the instantaneous amplitudes:

$$I_1 \cos(-\phi_1),\ I_2 \cos(-\phi_2),\ \ldots \text{etc.} \quad (62)$$

The various phase angles $\phi_1$, $\phi_2$, etc., then apply to this particular time origin. Thus the relative timing between the various waves is

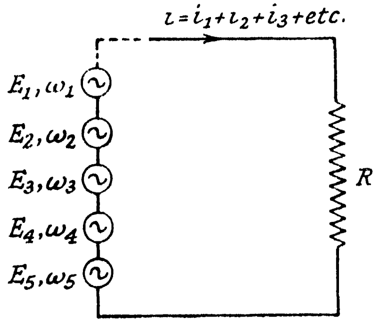



Fig. 16.—Synthesis of a Complex Current Wave by a Number of Generators in Series.

defined, and the form of the resultant complicated wave is fixed and determined completely by the expression 61.

The representation of two sinusoidal waves of equal frequency on the same vector diagram presents no difficulty, since the vectors are fixed relative to one another, and in the case of a vector diagram of the type in Fig. 11 they both rotate together at the angular speed $\omega$. If the vectors are to represent two sinusoidal waves of different frequencies, however, then they must be assumed to rotate relative to one another as well. Two such vectors are shown in Fig. 17 (a) at a certain instant of time. The resultant at this instant of time may be found by the usual parallelogram law, but as time varies this resultant will vary in length as well as in angular velocity. Its projection on to the real axis $ROR'$ at every instant of time will, however, represent the magnitude of the resultant current, since it is equal to the sum of the projections of the individual

vectors. Similarly for any number of sinusoidal component waves of different frequencies. But such a diagram is not of great use, and a better representation is given by using stationary vectors, giving the quantities in equation 62. These vectors are in such positions that they represent the various components at the instant $t=0$, and so the diagram will change as this chosen origin is varied, though for a given origin the vectors will all be stationary.

In Fig. 17 (b) the stationary vector diagram is shown, together with the conjugate vectors (dotted lines), so that the sum of the projections of all the vectors on to the axis $ROR'$ gives the amplitude of the resultant wave at $t=0$. The projections on the axis $SOS'$ cancel out at this and at every other instant of time. However

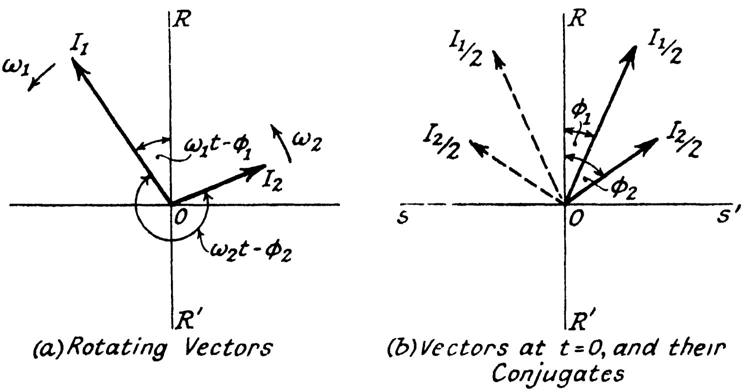



Fig. 17.—Vector Diagrams for Two Sinusoidal Waves of Different Frequencies. Similarly for any other number of waves of different frequencies.)

many component sinusoidal waves there are in the complete current $i$, they may all be represented by such pairs of stationary vectors, if the time origin be fixed, and the projections on to the axis $ROR'$ will represent the total magnitude of this current at $t=0$

If the value of $i$ at all other times is to be represented by such a vector diagram, the angular frequencies of every pair of vectors must be stated, but if these components are plotted in the form of a frequency spectrum diagram, all the information is provided that is necessary to define the complete wave and to plot its waveform. The spectrum of such a complex wave is shown in Fig. 18 (a), plotted in the same same way as the spectrum of the pure sinusoidal wave in Fig. 15 (a) and (b). The cosine terms, plotted as a symmetrical diagram about zero frequency, have amplitudes equal to

the projections of the vectors on to the axis *ROR′* (see Fig. 17 (b)), while the sine terms are plotted as a skew-symmetrical diagram, with amplitudes equal to the projections on to the other axis, *SOS′*. Then the sum of all these terms gives the complete wave of equation 61, with the time origin that has been chosen.

It has been shown in the introductory Sec. 3 that an electric wave of any form may be split up into a series (possibly infinite) of sinusoidal waves; we could consider such a wave to be synthesised by enough generators in series feeding a resistance load, in a circuit such as that shown by Fig. 16. The spectrum of the wave shows the magnitudes and phases of these sinusoidal components, which,

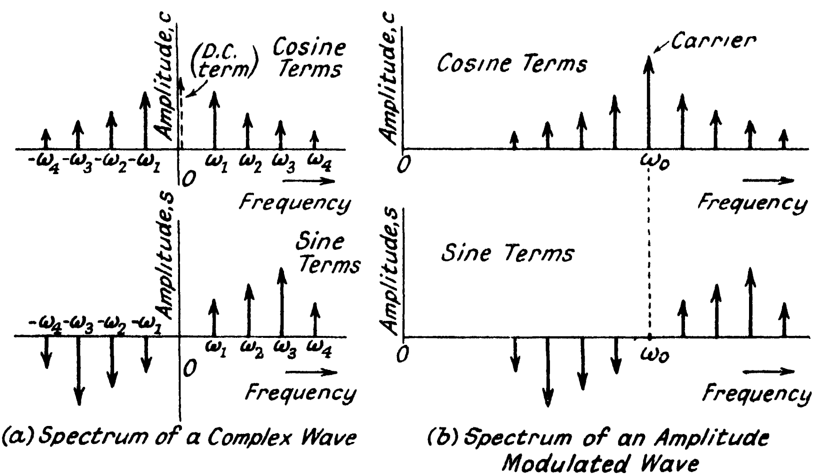



Fig. 18.—Spectra of Complex Waves Containing a Number of Sinusoidal Components.

when added together, compose the complete wave. In the particular way of drawing such spectra that has been described here, in terms of the cosine components C and the sine components S (at the same frequency), the amplitude and phase components are equal to:

$$\left.\begin{aligned} \text{amplitude} &= \sqrt{(C^2+S^2)} \\ \text{phase} &= \tan^{-1} S/C \end{aligned}\right\} \quad . \quad . \quad . \quad . \qquad (63)$$

There are certain special types of wave that we should consider—amplitude-modulated carrier waves, periodic non-sinusoidal waves, pulses of finite duration, etc.—together with their particular forms of spectra, since these may very conveniently be regarded as the

"standard forms" and other waveforms considered as combinations or modifications of these types. So far we have been discussing a general wave consisting of any number of sinusoidal terms of quite arbitrary amplitudes and phases, which may have any shape whatever. It is only when there is some definite relation between the various terms [3] that the resultant waveform takes on one of the "standard forms." Such relations may be between the amplitudes, the frequencies, or the phases of these various terms.

A wave of current has been spoken of specifically in this section, but naturally all the points concerning current waveforms and their geometry apply also to voltage waves; it is merely a matter of convenience.

## 14. The periodic wave—a Fourier series of harmonic terms

A number of different sinusoidal waves are said to be *harmonic* if the various frequencies are all multiples of some common frequency $\omega_1$, called the *fundamental (angular) frequency*. Thus the expression (61) for a series of waves becomes, in such a case:

$$i=I_1 \cos(\omega_1 t-\phi_1)+I_2 \cos(2\omega_1 t-\phi_2)+I_3 \cos(3\omega_1 t-\phi_3) + \ldots I_N \cos(N\omega_1 t-\phi_N) \ldots \text{etc.} \qquad (64)$$

Such a series of harmonic waves is the well-known *Fourier series*.[5] Certain harmonic terms may be absent, of course, but the above is the general harmonic series. The amplitudes $I_1, I_2 \ldots I_N$ and phase angles $\phi_1, \phi_2 \ldots \phi_N$ are still arbitrary, and the shape of the resulting waveform of $i$ may assume an indefinite number of different forms as these are varied, although the frequencies of the individual terms remain constant.

To express the shape of a particular waveform by such a harmonic series both the amplitudes and phases of all the terms must be specified.

Now suppose the vector diagram of Fig. 17 (a) represents two harmonic terms, so that, for instance $\omega_2=2\omega_1$. Then the vector $I_2$ will rotate twice while vector $I_1$ rotates once around the origin, and so after two cycles the vectors will be in the initial positions again. Similarly if we have any number of vectors rotating at angular velocities $\omega_1, 2\omega_1, 3\omega_1 \ldots N\omega_1$. If $N\omega_1$ is the highest, then the vectors will lie in identical positions after $N$ revolutions of this vector, which corresponds to one revolution of the fundamental $(\omega_1)$. Thus the projection of the resultant vector on to the axis $ROR'$, which represents the amplitude of the complex waveform,

will repeat itself at regular intervals, the length of the interval being a period of the fundamental term.

Such a wave, expressible as a harmonic series, is therefore called a *periodic wave*, its exact shape being determined by the various amplitudes and phases of the component harmonic terms.* Alternatively it may be more convenient to express the spectrum as cosine (even) and sine (odd) conjugate components, as we have already seen (Sec. 13). For instance, the spectrum in Fig. 18 (a) would represent a periodic wave if the components were uniformly spaced along the $\omega$ scale, that is if $\omega_1, \omega_2, \omega_3 \ldots$, etc., are made $\omega_1, 2\omega_1, 3\omega_1 \ldots$, etc. The repetition period of the wave depends on the fundamental angular frequency $\omega_1$:

$$\text{Frequency of fundamental } f_1 = \frac{\omega_1}{2\pi} \text{ cycles/sec.} \quad . \quad . \quad . \quad (65)$$

Hence:

$$\text{Fundamental period} = \frac{2\pi}{\omega_1} \text{ secs./cycle} \quad . \quad . \quad (66)$$

Fig. 19 (a) illustrates such a periodic (non-sinusoidal) wave, and in this case a D.C. or mean steady current is superimposed. This D.C. may be regarded as a wave of zero frequency and may be put on the spectrum diagram as two conjugate components, both together at zero frequency, the total length representing $I_0$, the steady D.C. value. These zero frequency components must be on the cosine spectrum, so that

$$\left.\begin{array}{ll} \text{when } \omega=0, & I_0 \cos \omega t = I_0 \\ \text{but} & I_0 \sin \omega t = 0 \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad (67)$$

So much, for the moment, on the *synthesis* of periodic complex waves from elementary sinusoidal terms; we shall come later to the *analysis* of a given complex wave into its component terms.

## 15. The amplitude-modulated carrier wave

A continuous sine wave is usually referred to as a *carrier wave* if it is used in a communication system for conveying a signal by some kind of modulation, or variation of its steady nature.

Suppose we express such a steady carrier wave as:

$$i = I_E \cos(\omega_0 t + \phi_0) \quad . \quad . \quad . \quad . \quad . \quad . \quad (68)$$

* See reference 7 for examples of addition of a few harmonic terms of different amplitudes and phases.

then we may use such a wave for conveying a signal by varying either $I_E$, $\omega_0$, or $\phi_0$, according to this signal. The resulting modulated wave is then called (in the three cases, respectively) an amplitude-, frequency-, or phase-modulated carrier wave. If more than one of these quantities is varied, the wave is called a *hybrid wave.* The result of modulation is that the simple spectrum of the steady carrier wave (angular frequency $\omega_0$) spreads out to cover a range of frequencies, and the sinusoidal terms in this spectrum, which when added together constitute the modulated wave, are called the *sidebands.*

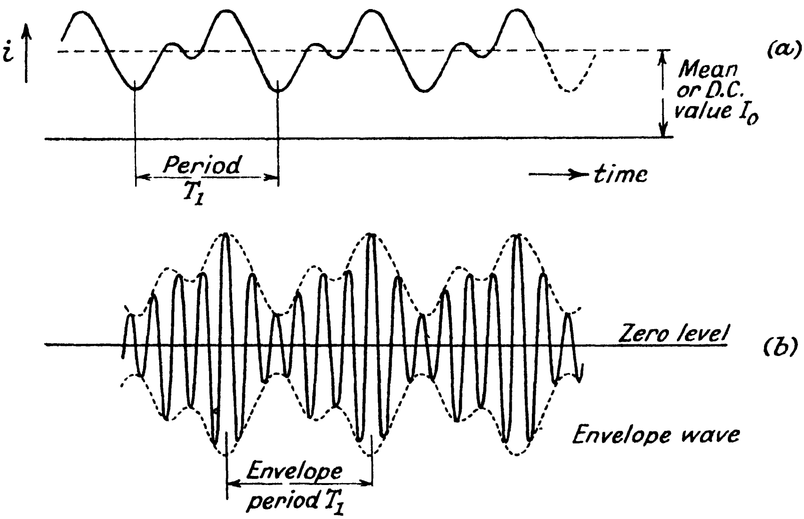



Fig. 19.—Periodic Complex Waves. (a) An Envelope Wave with a Positive D.C. Component. (b) A Carrier Wave Modulated in Amplitude by the same Envelope Wave (a).

The most important type of modulation, from our point of view here, is amplitude modulation,[6] so let us take the simple case of a carrier wave being varied in amplitude sinusoidally. The varying amplitude of the wave (called the wave envelope*) is then:

$$I_E = mI_0 \cos(\omega t - \phi) \qquad (69)$$

The amplitude-modulated wave is then:

$$i = I_0[1 + m\cos(\omega t - \phi)] \, . \cos \omega_0 t \qquad (70)$$

if we choose the time origin, for simplicity, to pass through a peak of the carrier wave, so that $\phi_0 = 0$ (or $\eta\pi$).

* A more exact definition of "envelope" is given in Sec. 67.

The variations in amplitude are superimposed on a steady or average carrier amplitude $I_0$, which accounts for the addition of the D.C. term to the envelope of this modulated wave. The constant $m$, called the *modulation index*, fixes the fractional relation* between the envelope and steady carrier amplitudes, as is illustrated in Fig. 20 (b).

Only if there is some harmonic relation between the carrier and envelope angular frequencies, $\omega_0$ and $\omega$, will the modulated wave remain stationary with respect to its envelope. Otherwise, when such a wave was being viewed on an oscillograph with the time-base synchronised to the envelope frequency, the actual modulated

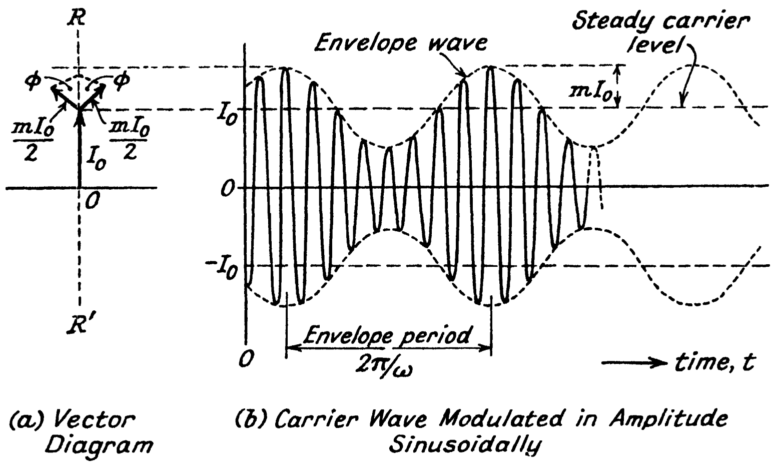



Fig. 20.—The Amplitude-modulated Carrier Wave (see equation 70); Envelope Wave$=mI_0 \cos(\omega t+\phi)$; Depth of Modulation$=m$.

wave would in general move about, giving a blurred image, although its outline (the envelope) would remain stationary and clear.

Before considering the vector diagram, let us expand the wave expression 70 into its sinusoidal terms, or sidebands. This equation may be written:

$$i=[I_0 \cos \omega_0 t]+\left[\frac{mI_0}{2}\cos(\overline{\omega_0+\omega t}-\phi)\right]+\left[\frac{mI_0}{2}\cos(\overline{\omega_0-\omega t}+\phi)\right] \quad \ldots \ldots \ldots \ldots (71)$$

which shows that such a wave consists of a steady term of frequency $\omega_0$, the carrier component, and steady terms of frequencies $(\omega_0+\omega)$,

* Usually called the "depth of modulation" or, as a percentage of $I_0$, the "percentage modulation."

the upper sideband, and $(\omega_0 - \omega)$, the lower sideband. That is, a steady carrier wave with two other waves, one higher in frequency by an amount equal to the envelope frequency and one lower in frequency by the same amount.

This is the essential property of the spectra of amplitude-modulated waves— such spectra have exact symmetry about a central term. But we have already been recommending, in previous sections, that all spectra be plotted in terms of conjugate components, which automatically makes them of a symmetrical form about zero frequency, so that we have to be careful. The symmetry that is required of a spectrum to make it give an amplitude-modulated wave is symmetry about a central term of frequency $\omega_0$, *which is not zero*. Before going further into this important point, a word about the vector diagram of Fig. 20 (a) is needed.

This is plotted in a conventional way, the carrier component vector $I_0$ along the vertical axis $ROR'$ and the two sideband vectors of length $mI_0/2$ rotating in opposite directions about the vector $I_0$, at angular speeds $\pm\omega$. At time $t=0$, the instantaneous positions of the vectors gives the amplitude of the modulated wave at $t=0$, the phase angles $\pm\phi$, being between the sideband vectors and the axis $ROR'$. Now if we imagine that the whole diagram rotates at the angular speed $\omega_0$, then this diagram represents the complete modulated carrier wave by its resultant projection on to the axis $ROR'$. But such an idea is of little use, since it is difficult to imagine vectors rotating at different speeds, so we can discard this as we did in Sec. 13 when considering waves of unrelated frequencies and take the vector diagram as shown in Fig. 20 (a). This diagram then is the result of reducing the whole angular speed of the vectors by the carrier angular frequency $\omega_0$, and the sideband vectors which rotate at speeds $\pm\omega$ are shown, on this stationary vector diagram, in their positions corresponding to $t=0$.

The projection of the vectors, in such a diagram, on to the axis $ROR'$ then describes the wave *envelope*, added on to its steady average level equal to the carrier component $I_0$. From equation 70 this complete envelope curve is:

$$i_E = I_0[1 + m\cos(\omega t - \phi)] \quad . \quad . \quad . \quad . \quad (72)$$

The truth of this statement is readily verified from Fig. 20 (a) and (b). Now this same complete envelope would be described by this diagram whatever the carrier frequency, since we have drawn the sideband vectors as being relative to the carrier at an angular

velocity $\omega_0$, and so all trace of the carrier frequency is omitted. This suggests that we may transfer the carrier frequency to zero, the same vector diagram being applicable, and that a great simplification may be obtained, when we are dealing with amplitude modulated waves, by considering the envelope only and forgetting the carrier. The envelope wave is a simpler kind of wave than a modulated carrier, and any results that we may obtain in practice, for instance concerning distortion of the envelope, may be applied directly to the modulated wave, whatever its carrier frequency.

This simple theorem is of great importance and we should consider it in more detail.

## 16. Transferring the carrier frequency to zero

Continuing with the suggestion made in the preceding section, let us apply the idea of transferring the carrier frequency to other frequencies, and even to zero, in the spectrum of such a wave.

The simple spectrum consists of three terms, as given by equation 71, and we could if we wished split these terms into their cosine and sine components and plot them as cosine and sine spectra, just as we did in equation 60 for the simple sinusoidal wave. This would result in a spectrum of three cosine and three sine terms at frequencies $\omega_0$, $(\omega_0+\omega)$, and $(\omega_0-\omega)$ with a conjugate spectrum of three terms at frequencies $-\omega_0$, $-(\omega_0+\omega)$, and $-(\omega_0-\omega)$. The reader may expand equation 71 to determine the exact expression for this complete spectrum, and the result will be that shown in Fig. 21 (a). But this is unnecessary, since we may again use the idea given in Sec. 15 of transferring the carrier frequency to zero.

If we derive the cosine and sine components of the vectors as plotted in Fig. 20 (a), in which the carrier vector has been assumed to be stationary, we obtain:

$$i=I_0+\left(\frac{mI_0}{2}\cos\phi\right)\cos\omega t+\left(\frac{mI_0}{2}\cos\phi\right)\cos(-\omega t)$$
$$+\left(\frac{mI_0}{2}\sin\phi\right)\sin\omega t-\left(\frac{mI_0}{2}\sin\phi\right)\sin(-\omega t) \quad . \quad . \quad . \quad . \quad (73)$$

These components may be compared with those for a simple sinusoidal wave as given by equation 60. If we sum them we obtain the expression 72 for the complete envelope wave $i_E$.

These components are plotted in Fig. 21 (b) as a spectrum symmetrical about zero frequency, which is seen to be identical with the spectrum of a simple sinusoidal wave as in Fig. 15 (a) and (b).

One important point—the lengths of the components in Fig. 21 (b), when the carrier has been "transferred to zero," are **C** and **S**, where

$$\mathbf{C}=\frac{mI_0}{2}\cos\phi$$
$$\mathbf{S}=\frac{mI_0}{2}\sin\phi \tag{74}$$

which are twice the lengths of the components in the complete spectrum for the modulated wave of frequency $\omega_0$, as shown in Fig. 21 (a).

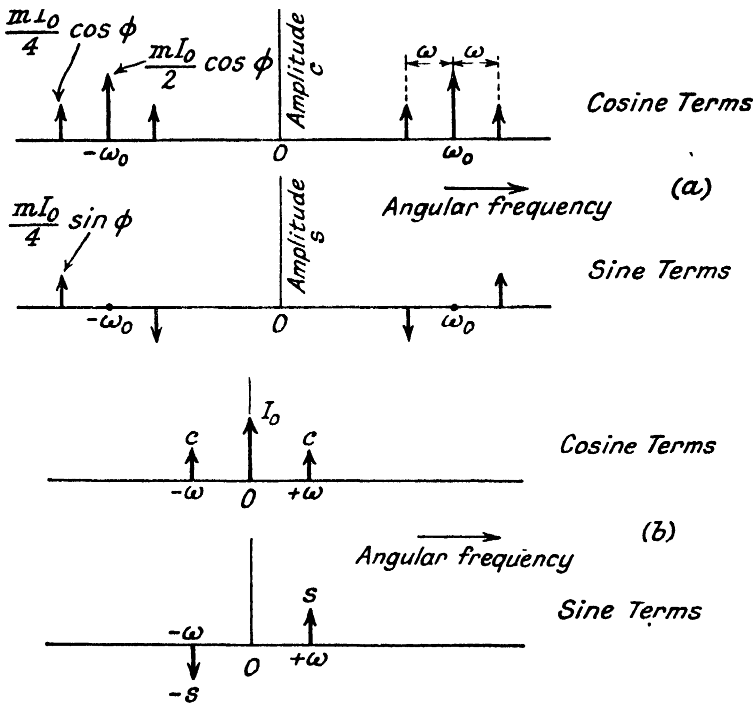



Fig. 21.—Illustrating the Theorem Concerning the Transfer of a Carrier to Zero Frequency.

(a) Complete Conjugate Spectrum of a Carrier Wave Modulated in Amplitude by a Sinusoidal Envelope.
(b) Spectrum of this Modulated Wave, with the Carrier Frequency Transferred to Zero, which is Identical with the Spectrum of the Envelope.

This theorem may obviously be extended to the case of a carrier wave being modulated in amplitude by an envelope of any complex

waveform. For instance, take a periodic wave with a steady D.C. component such as that shown in Fig. 19 (a). Let us suppose this wave is the result of adding together the Fourier series of harmonic waves given in equation 64, the various amplitudes $I_1$, $I_2$, $I_3$ . . . and phases $\phi_1, \phi_2, \phi_3$ . . . being fixed. Then the carrier wave shown in Fig. 19 (b) modulated by this same envelope wave will be:

$$i=I_0\,[1+m_1\cos(\omega_1 t-\phi_1)+m_2\cos(2\omega_1 t-\phi_2)+\ldots]\cos\omega_0 t \qquad (75)$$

if we write:

$$I_1=m_1 I_0,\quad I_2=m_2 I_0,\ \text{etc.} \quad \ldots\ldots \quad (76)$$

Expanding this modulated wave into its carrier and sideband components gives:

$$i=I_0\cos\omega_0 t+\frac{m_1 I_0}{2}\cos(\overline{\omega_0+\omega_1}\,t-\phi_1)+\frac{m_2 I_0}{2}\cos(\overline{\omega_0+2\omega_1}\,t-\phi_2)$$
$$+\frac{m_1 I_0}{2}\cos(\overline{\omega_0-\omega_1}\,t+\phi_1)+\frac{m_2 I_0}{2}\cos(\overline{\omega_0-2\omega_1}\,t+\phi_2)\ .\ .\ . \qquad (77)$$

which is a carrier term (angular frequency $\omega_0$) and a series of pairs of sidebands symmetrically disposed about this carrier, there being one pair of sidebands for each sinusoidal component in the envelope.

The spectrum diagram for such a wave is therefore similar to that for the carrier wave with a pure sinusoidal envelope (see Fig. 21) except that there are many pairs of sideband terms. Each term in the above equation may be split into its cosine and sine terms, for example:

$$\frac{m_1 I_0}{2}\cos(\omega_0+\omega_1 t-\phi_1)=\left[\frac{m_1 I_0}{2}\cos\phi_1\right]\cos(\omega_0+\omega_1)t$$
$$+\left[\frac{m_1 I_0}{2}\sin\phi_1\right]\sin(\omega_0+\omega_1)t \quad .\ . \quad (78)$$

and the spectrum may be plotted as sets of conjugate terms, as previously (e.g. Fig. 21 (a)).

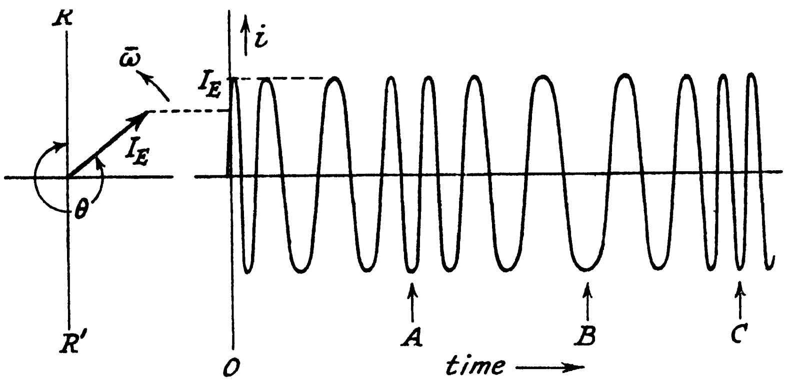
However, without further explanation, we may apply the theorem concerning the transfer of the carrier frequency to zero, which affords such a simplification. The spectrum is shown in Fig. 18 (b), plotted conventionally about the central carrier term of angular frequency $\omega_0$ (the corresponding conjugate terms central about the frequency $-\omega_0$ are omitted here) and if the carrier frequency be "transferred to zero" the spectrum of Fig. 18 (a) is the result.

But such a spectrum has already been referred to as the spectrum of the envelope wave (64); thus the spectrum of an amplitude

modulated wave is identical with the spectrum of its envelope, except that its axis of symmetry is not at zero but at the carrier frequency $\omega_0$. It is clear that this theorem will be true whatever the form of the envelope, since it may be applied to every pair of sideband terms in the modulated wave spectrum.

## 17. Frequency-modulated and phase-modulated carrier waves—some definitions and distinctions.

It will be as well to discuss here briefly some of the characteristics of the waveforms of frequency-modulated and phase-modulated* waves, since these types of modulation have come into common use during recent years.[8, 9]




(a) The Vector (b) The Frequency- or Phase-Modulated Wave

Fig. 22.—Vector Rotating at a Varying Speed Represents a Frequency- or Phase-modulated Wave of Constant Amplitude.

If we take expression 68 as the general carrier wave, we may then modulate this wave according to the signal to be transmitted, by varying $\omega_0$ or $\phi_0$. The amplitude $I_E$ of the modulated wave is constant in this case and the wave presents the general appearance of Fig. 22 (b). From the general waveform appearance it is not possible to distinguish between these two types of modulated wave (except in certain cases, as we shall see).

Such a wave may be represented by the rotating vector $I_E$, of constant length as in Fig. 22 (a) which, however, does not rotate at a constant rate $\omega_0$ radians/sec., since such a vector would represent

* Often written as F.M. and P.M. (as opposed to A.M., for Amplitude-modulated).

a constant-frequency carrier wave (as in Fig. 11). It is most usual, in practical communication systems, to use a carrier wave having an average frequency value $\omega_0$ and to modulate the frequency on either side of this.

To make the relations clearer between frequency modulation and phase modulation let us consider the simple case of conveying the signal:

$$e = E_m \cos \omega_m t \quad \ldots \ldots \quad (79)$$

by these two methods.

First, in the case of frequency modulation the angular velocity of this vector (Fig. 22 (a)) must be varied according to the signal, so that if $\hat{\omega}$ represents its velocity at any instant

$$\hat{\omega} = \omega_0(1 + k \cos \omega_m t) \quad \ldots \ldots \quad (80)$$

where $k\omega_0$ is the *deviation*, that is the shift of the carrier frequency corresponding to the peak signal $E_m$. This term $\hat{\omega}$ is called the *instantaneous frequency* of the modulated wave.

Such a variation of frequency may be produced, for example, by varying the capacity in a tuned circuit by the use of special valve modulators, [17, 18] so that the frequency of tuning changes according to the voltage applied to the valve grid. It would be out of place to discuss such methods here, but the reader may find many descriptions of practical techniques in the literature.

Now if the vector in Fig. 22 (a) rotates at the angular speed $\hat{\omega}$ and if $\theta$ is the angle between the vector and the reference axis *ROR'* at any instant, then

$$\hat{\omega} = d\theta/dt \quad \ldots \ldots \quad (81)$$

and we may write the wave represented by the vector's projection on *ROR'* as:

$$i = I_E \cos \theta \quad \ldots \ldots \quad (82)$$

Then the frequency-modulated wave is given by:

$$i = I_E \cos \left[ \omega_0 t + \frac{k\omega_0}{\omega_m} \sin \omega_m t \right] \quad \ldots \quad (83)$$

(from equation 80).

$$\text{The coefficient } \frac{k\omega_0}{\omega_m} = \frac{\text{Peak shift of the carrier frequency}}{\text{Modulating frequency}}$$

and is called the *modulation index*, or *deviation ratio*.

Secondly, phase modulation. Suppose we have a means of varying the phase angle $\phi_0$ of the steady carrier wave (68) according

to the signal $e = E_m \cos \omega_m t$. Let the peak phase-shift (called the *phase deviation*) be $k'\phi_0$, corresponding to the peak signal $E_m$. The phase-modulated carrier wave is then given by:

$$i = I_E \cos (\omega_0 t + k'\phi_0 \cos \omega_m t) \quad . \quad . \quad . \quad (84)$$

since the angle $\theta$ in this case is $\omega_0 t$ plus a sinusoidal phase variation. If these two expressions (83, 84) be compared it may be seen that the carrier wave in both cases has a *steady* (average) angular frequency $\omega_0$ with a phase angle which varies depending on the signal. Thus we may consider that frequency and phase modulation lead to *similar* results as regards the actual transmitted waveform. This point is made clearer by the diagram of Fig. 22, in the following way.

At the point $A$ on this diagram the "instantaneous frequency" of the wave is seen to be higher than at the point $B$. Also at the point $C$ the "instantaneous frequency" is the same as at $A$. But the phase at $C$ may be different from that at $A$, as measured in relation to a steady unmodulated wave of this frequency. Then we may regard this wave as a frequency-modulated wave, since its frequency is changing from $A$ to $B$ and back again at $C$. Alternatively, we may regard it as having been modulated in phase from $A$ to $C$, but in so doing *its frequency must have been changed, or modulated, between these points.*

## 18. Frequency- and phase-modulated waves—the steady-state vector diagram

The frequency- and phase-modulated waves have a steady, constant amplitude but are not purely sinusoidal in waveform; thus they may be divided up into a number of steady sinusoidal components, or sidebands. The spectrum of sideband components will depend on the particular form of the modulating wave, and it may be shown to possess *always* an infinite number of components. The actual form of the sideband spectrum may be made clear by considering the vector diagram, which may be drawn in a manner similar to that adopted for the amplitude-modulated wave.[9, 10]

In Fig. 20 (a) there is shown the vector diagram of an amplitude-modulated wave (sinusoidal envelope), with the carrier component $I_0$ stationary on the paper; the carrier frequency has been "reduced" to zero frequency. The two sidebands rotate in opposite directions about the carrier component at the modulation angular frequency. Now suppose we turn these two sideband vectors through 90°, as

in Fig. 23 (a), we obtain an approximation to the vector diagram of a frequency- or phase-modulated wave.[11] The stationary vector $OP$ is the steady carrier component, of frequency $\omega_0$, and $PQ$, $PQ'$ two sideband components rotating in opposite directions about $P$. The resultant is then a vector which oscillates from side to side, as in Fig. 23 (b). This would represent a frequency- or phase-modulated wave with the mean angular frequency $\omega_0$, except for the fact that this resultant vector varies in length also, from $OA$ to $OA'$. This variation in length may be partially corrected by the further addition of a pair of vectors $QN$ and $Q'N'$, symmetrical about the axis $ROR'$, as shown in Fig. 23 (c). It is clear that for an exact representation of a frequency- or phase-modulated wave,

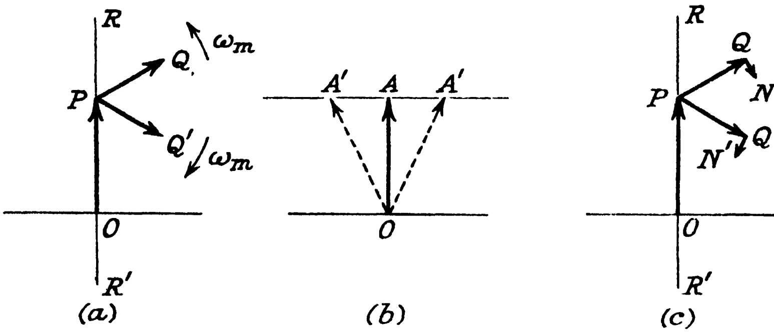



Fig. 23.—Frequency- and Phase-modulated Waves. Carrier and Sideband Vectors.

in this way, an infinite number of such pairs of sideband vectors will be needed.

If the modulation index be small, then the vector in Fig. 23 (b) will swing about $OA$ by a small angle only, and consequently the change in length $OA$ to $OA'$ will be small; thus the approximation to the wave is less inexact should a few sideband component pairs be used.

## 19. The spectra of frequency- and phase-modulated carrier waves

Exact calculation of the lengths of the vectors $OP$, $PQ$, $QN$, etc. in the general case is very difficult, and essentially a problem in pure mathematics. We shall give the barest outline here of the method used, but references are included at the end of the chapter for those readers particularly interested in this matter.[8, 10, 12, 13] Fortunately

it is rare that exact calculation of these spectra terms needs to be made, though a general idea of their forms is essential.

Unfortunately there is no theorem concerning the transfer of the carrier to zero frequency for these particular types of modulation, as there was for amplitude modulation (Sec. 16). The spectrum of the modulated carrier wave is not so simply related to the spectrum of the modulating wave as to make this possible.[15]

The phase-modulated wave (equation 84) may be divided into two orthogonal components, since it is of the form $\cos(x+y)$. The in-phase or cosine component has an amplitude:

$$\mathrm{C}=I_E \cos(k'\phi_0 \,.\, \cos \omega_m t) \quad . \quad . \quad . \quad . \quad (85)$$

and the quadrature or sine component:

$$\mathrm{S}=-I_E \sin(k'\phi_0 \,.\, \cos \omega_m t) \quad . \quad . \quad . \quad (86)$$

The frequency of both components is constant, at $\omega_0$. Similar expressions may be found for the frequency-modulated wave (equation 83). Now these expressions involve a cosine of a cosine and a sine of a cosine, and these may be expanded into an infinite series of Bessel functions*; in any particular example these Bessel functions may be found in lists of tables and they are the amplitudes of the sideband components. The frequencies of these components lie symmetrically about the carrier frequency $\omega_0$ by intervals equal to the modulating frequency and all its harmonic frequencies. In the case of modulation by a complex wave, the sidebands are spaced from the carrier frequency by each of the modulating frequencies $\omega_n$, $\omega_m$, $\omega_l$ . . ., etc., and their harmonics, and there are also components at every sum and difference frequency $\omega_n \pm \omega_m$, $\omega_n \pm \omega_m \pm \omega_l$ . . ., etc., and the spectrum becomes extremely complex.[13, 15]

Examples of simple F.M. spectra, with sinusoidal modulation are illustrated in Fig. 24 (a) for $k\omega_0/\omega_m=16$ and (b) for $k\omega_0/\omega_m=4$. The general differences should be noted; the sidebands are spaced at intervals equal to the modulation frequency $\omega_m/2\pi$, and there are more reversals of the sideband amplitudes with the higher index of 16. Another important point is that these spectra spread outside the deviation frequencies and that the spread is more, as may be expected, in the case of the higher frequency modulation. Obviously if the carrier be modulated in frequency *very* gradually there will be no components of appreciable amplitude outside the range of modulation.

* See reference 14.

There is one important point in connection with the spectra of F.M. and P.M. waves, which they have in common with A.M. waves. These spectra forms are independent of the carrier frequency provided that the modulation indices are constant, though, as was mentioned earlier, they are not identical with the spectra of the modulating waveforms.

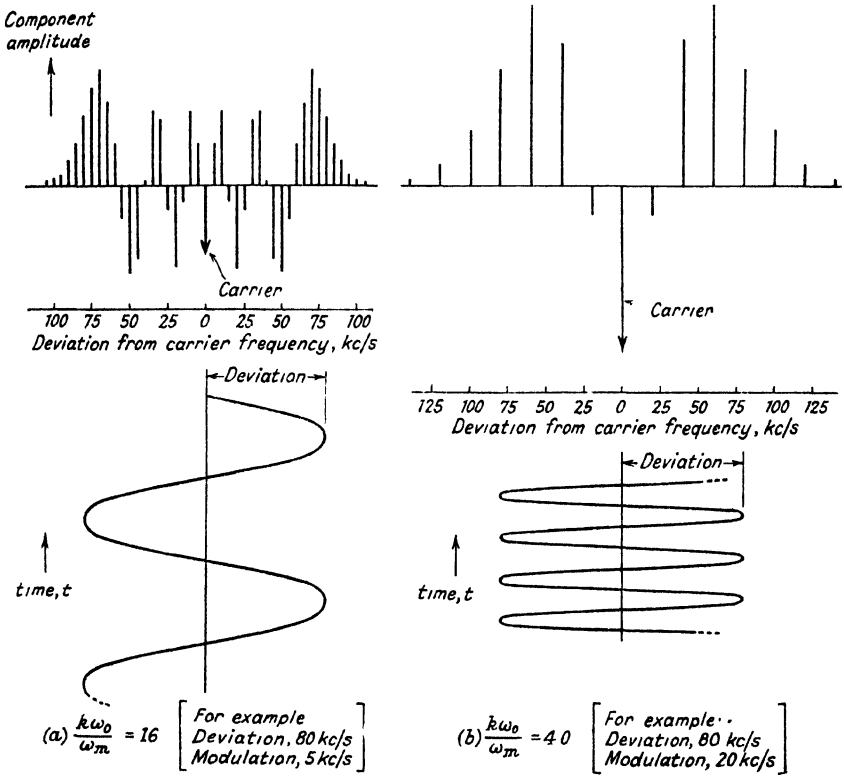



Fig. 24.—Spectra of F.M. Waves with Different Deviation Ratios.

## 20. Fourier series analysis of a periodic wave

We have dealt in the preceding sections with the addition of sinusoidal components of different frequencies, and have shown the characteristically different types of wave formed when the various components have special relations between their amplitudes and phases. In particular, it was shown in Sec. 14 that the addition of a series of harmonic components resulted in a periodic non-sinusoidal wave.

How may we perform the reverse process and discover the amplitudes and phases of the constituent sinusoidal waves when the resultant waveform is known? For example, Fig. 25 (a) shows a "square wave," a waveform commonly used in network transient calculation and testing, being the waveform of an E.M.F. which reverses in sign at regular time intervals $T_0/2$ $(=T_1)$. Such a wave is periodic and so we may start with the assumption that it may be represented by the sum of a harmonic series of cosine or sine waves, or both. Let the series be:

$$e=a_0+\sum_{n-1}^{\infty} a_n \cos\left(\frac{2\pi n\,.\,t}{T_0}\right)+\sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi n\,.\,t)}{T_0}\right) \quad . \quad . \quad . \quad (87)$$

where $a_0$ is the average or direct current value and $a_n$, $b_n$, the unknown amplitudes of the harmonic components. This series is similar* to the harmonic series, equation 64. Note that the fundamental angular frequency is now written $2\pi/T_0$ instead of $\omega_1$; thus $\omega_1$ is now expressed in terms of the fundamental period of repetition $T_0$ (see equation 66).

By choice of the time origin this series (87) may be simplified. With the origin at the point (1) on Fig. 25 (a) the wave is symmetrical about $t=0$, being a mirror image, and in this case a *cosine series only* is needed. Since a cosine wave, $A \cos \omega t$, is symmetrical in shape about $t=0$, it follows that the sum of any number of cosine waves is symmetrical too. Also, with the origin at the point (2), a *sine series only* is needed since the wave is skew-symmetrical (an inverted mirror image) about $t=0$. Before analysing any wave into its Fourier components in a practical example, the origin should be chosen carefully, to save labour.

For this example, let us take the origin at the point (1), so that the amplitudes $b_n$ of all sine terms in the series 87 are zero. The amplitudes $a_n$ of the cosine components in this Fourier series are found as follows:

Multiply the expression for the wave by cos $(2\pi r.t/T_0)$ where $r$ may then be given the values 1, 2, 3 . . ., etc., in turn; then integrate the product over a complete fundamental period $T_0$. To make the reason for this process clearer, let us write the wave $e$

* These equations represent the two trigonometrical forms of a Fourier Series; equation 64 gives the amplitudes and phases of the harmonics and equation 87 gives the cosine and sine components. Their relation is given by equation 63.

as $f(t)$—that is the square wave in our present example. Then this integral becomes (from 87), after putting $n=1, 2, 3$, etc.:

$$\int_{-T_0/2}^{+T_0/2} f(t)\cos\left(\frac{2\pi r.t}{T_0}\right)dt=\int_{-T_0/2}^{+T_0/2}\left[a_0\cos\left(\frac{2\pi r.t}{T_0}\right)+a_1\cos\left(\frac{2\pi t}{T_0}\right).\cos\left(\frac{2\pi r.t}{T_0}\right)\right.$$
$$+a_2\cos\left(\frac{4\pi t}{T_0}\right).\cos\left(\frac{2\pi r.t}{T_0}\right)$$
$$\left.+\ldots\right]dt \quad . \quad . \quad . \quad (88)$$

Now we may give $r$ the values 1, 2, 3, etc., in turn and carry out these integrations. It will be found, however, that the integral on the right-hand side vanishes except for the one term for which $r=n$, that is for the particular harmonic whose frequency is equal to the frequency of the multiplying wave $\cos(2\pi r.t/T_0)$. There is an example of such an integration with which the reader may be familiar—the power delivered by a sinusoidal current $i$ under a certain E.M.F., $e$, also of sinusoidal waveform, is a constant ($ei$) only if the frequencies are equal. In the present example we are multiplying a series of harmonic sinusoidal terms by another single sinusoidal term whose frequency we may vary in steps to be equal to the frequencies of the various terms in the series; then we are integrating the product so that this process may be likened to finding the power. The power, as given by the product of these two waves, one complex and the other sinusoidal, is given entirely by the product of the sinusoidal terms of equal frequencies.

Mathematically, this is given by the identities:

$$\int_{-T_0/2}^{+T_0/2} a_0\cos\left(\frac{2\pi r.t}{T_0}\right)dt=0, \qquad \int_{-T_0/2}^{+T_0/2} a_n\cos\left(\frac{2\pi n.t}{T_0}\right)\cos\left(\frac{2\pi r.t}{T_0}\right)dt=0$$

$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad (89)$$

also it should be noted $\displaystyle\int_{-T_0/2}^{+T_0/2} b_n\sin\left(\frac{2\pi n.t}{T_0}\right)\sin\left(\frac{2\pi r.t}{T_0}\right)dt=0$

unless $r=n$, in which case the product integrals equal $a_nT_0/2$ or $b_nT_0/2$ respectively.

Another product integral should be noted:

$$\int_{-T_0/2}^{+T_0/2} a_n \cos\left(\frac{2\pi n \cdot t}{T_0}\right) \cdot \sin\left(\frac{2\pi r \cdot t}{T_0}\right) dt = 0 \quad \text{always} \quad . \quad . \qquad (90)$$

and also if $n$ and $r$ are interchanged, since they are both integers.

For our present example of a square wave (Fig. 25 (a)) with the time origin at the point (1), we have:

$f(t)=E$ between the limits $t=-\frac{T_0}{4}$ to $+\frac{T_0}{4}$

and $f(t)=0$ between the limits $-\frac{T_0}{2}$ to $-\frac{T_0}{4}$ and also between $+\frac{T_0}{4}$ and $+\frac{T_0}{2}$.

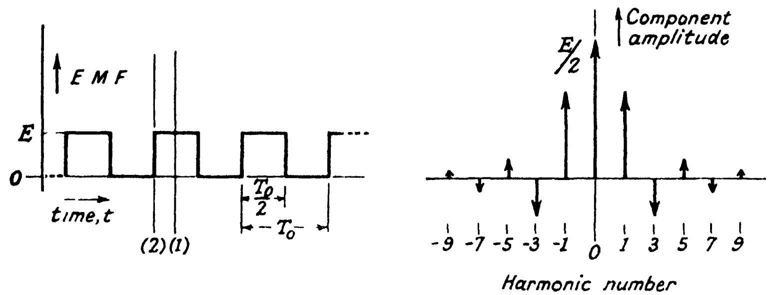



Fig. 25.—A Continuous Square Wave and its Harmonic Spectrum.
(a) Continuous Square Wave—origin chosen at (1) for spectrum analysis.
(b) Spectrum (conjugate terms) of Continuous Square Wave (cosine terms only).

Then equation 88 becomes, after substituting for $f(t)$:

$$\int_{-T_0/4}^{+T_0/4} E \cos\left(\frac{2\pi n \cdot t}{T_0}\right) dt = a_n \frac{T_0}{2} \quad . \quad . \quad . \quad . \qquad (91)$$

all other terms being zero. Carrying out this integration and putting in these limits:

$$\frac{2E}{n\pi} \cdot \sin\left(\frac{n\pi}{2}\right) = a_n \quad . \quad . \quad . \quad . \quad . \quad . \qquad (92)$$

which gives the value of $a_1, a_2, a_3$ . . ., etc., if we put $n=1, 2, 3$ . . ., etc. However, the term $\sin \frac{n\pi}{2}$ becomes zero or unity according as $n$ is even or odd.

This spectrum is illustrated by Fig. 25 (b), and is seen to consist of a series of harmonics, alternately positive and negative, plotted on an arbitrary horizontal scale in terms of $n$. In this case the harmonics are zero for $n=2, 4, 6, 8$ . . ., etc.: that is, the spectrum consists of *odd harmonics only*. The terms are plotted as conjugates (see Sec. 13), and are *cosine*.

If we had chosen the time origin in this diagram at the point (2), the amplitudes of the cosine terms, $a_n$, would all have been zero, and we could have worked out the values of the sine terms, $b_n$, by a process similar to the one described above. Thus multiplying by $\sin 2\pi n \,.\, t/T_0$ and integrating over a whole period $T_0$ would give the following equation:

$$\int_0^{+T_0/2} E \sin \frac{2\pi n}{T_0} t \,.\, dt = b_n \frac{T_0}{2} \quad . \quad . \quad . \quad . \quad (93)$$

similar to 91, for the amplitudes $b_n$ of the sine components in the square wave.

We have not yet found the value of $a_0$, the steady or D.C. component of this wave. This square wave clearly has a D.C. component since it is unidirectional (see Fig. 25 (a)). This component, $a_0$, is the average value of the wave and may be found by the same means as that used for finding the mean ordinate of any graph—by finding the area and dividing by the base length. Integrating this waveform over a single period is sufficient, since it is periodic.

$$a_0 = \text{D.C., or average value} = \frac{1}{T_0} \int_{-T_0/2}^{+T_0/2} f(t) \,.\, dt \quad . \quad . \quad (94)$$

in the general case; in our particular example this average value $a_0$ clearly equals $E/2$. We may consider this component as being the amplitude of a cosine term of zero frequency, since

$$a_0 \cos(0) = a_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (95)$$

and it is convenient to plot $a_0$ on the spectrum as a *zero order harmonic*, i.e. at $n=0$, as has been done in the diagram, Fig. 25 (b).

To summarise, we may express a periodic complex waveform $f(t)$ as a harmonic series:

$$f(t) = \sum_{n=0}^{\infty} \left. \begin{matrix} a_n \cos \\ b_n \sin \end{matrix} \right\} \frac{2\pi n \,.\, t}{T_0} \quad . \quad . \quad . \quad . \quad (96)$$

where the amplitude coefficients $a_n$ and $b_n$ are given by:

$$\left.\begin{matrix} a_n \\ b_n \end{matrix}\right\} = \frac{2}{T_0}\int_{-T_0/2}^{+T_0/2} f(t) \left.\begin{matrix} \cos \\ \sin \end{matrix}\right\} \frac{2\pi n . t}{T_0} . dt \quad . \quad . \quad . \quad . \quad (97)$$

the choice of the time origin deciding whether cosine or sine terms, or both, are required.

If the wave to be analysed $f(t)$ is neither quite symmetrical nor skew-symmetrical about the chosen time origin, as in the example

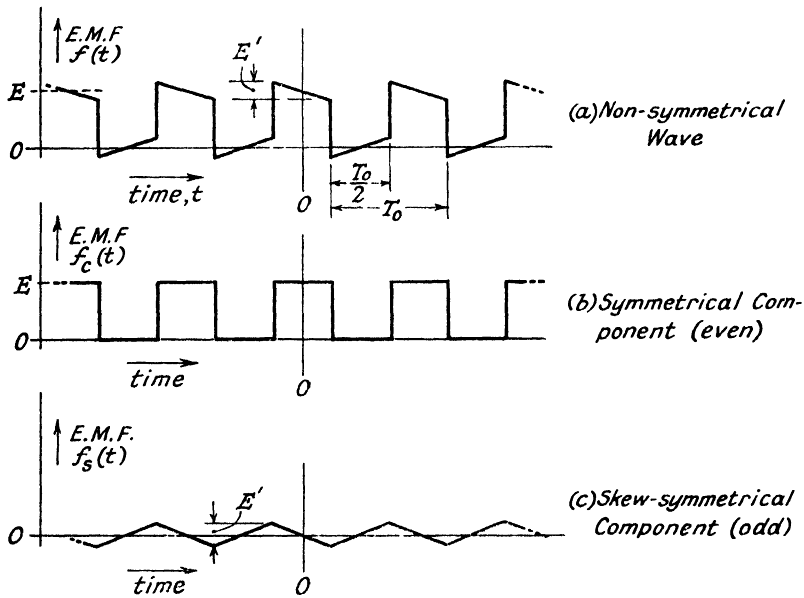



Fig. 26.—Division of a Non-symmetrical Wave into Odd and Even Components.

of Fig. 26 (a), then a series of both cosine and sine harmonic terms must be found, as equation 87. The cosine series alone will represent an even or symmetrical function $f_c(t)$ and the sine series alone an odd or skew-symmetrical function $f_s(t)$; the wave may, if desired, first be split up into these two components. For this example these odd and even components are illustrated in the same Figure, 26 (b) and (c). These components may always be found in any practical example by plotting the non-symmetrical wave $f(t)$ and then plotting the wave backwards (with the time scale reversed) $f(-t)$. The even component is then given by adding corresponding

ordinates, as $f_c(t)=f(t)+f(-t)$, and the odd component by subtracting corresponding ordinates, as $f_s(t)=f(t)-f(-t)$.

Each series may then be found separately by multiplying the even component $f_c(t)$ by cos $2\pi nt/T_0$ and integrating over a period $T_0$, and the odd component $f_s(t)$ by sin $2\pi nt/T_0$ and integrating. Exactly the same result is obtained of course by multiplying the expression *for the whole wave* $f(t)$ (Fig. 26 (a)) by these cosine and sine functions separately, and integrating.

The spectra of cosine and sine terms may be plotted separately, as in Fig. 18 (a); if, as in that diagram, conjugate terms are used the cosine spectrum will be symmetrical and the sine spectrum skew-symmetrical about zero frequency, as we saw in Sec. 13, since:

$$\left.\begin{aligned} a_n \cos\left(\frac{2\pi n\,.\,t}{T_0}\right) &= \frac{a_n}{2}\cos\left(\frac{2\pi n\,.\,t}{T_0}\right)+\frac{a_n}{2}\cos\left(-\frac{2\pi n\,.\,t}{T_0}\right) \\ \text{and } b_n \sin\left(\frac{2\pi n\,.\,t}{T_0}\right) &= \frac{b_n}{2}\sin\left(\frac{2\pi n\,.\,t}{T_0}\right)-\frac{b_n}{2}\sin\left(-\frac{2\pi n\,.\,t}{T_0}\right) \end{aligned}\right\}. \quad (98)$$

as in equation 58. (See also Fig. 13.)

There are two other methods of writing the Fourier series of equation 96, for a non-sinusoidal wave, which should be mentioned. First, the orthogonal components of a single-frequency term may be combined thus:

$$a_n \cos\left(\frac{2\pi n\,.\,t}{T_0}\right)+b_n \sin\left(\frac{2\pi n\,.\,t}{T_0}\right)=\sqrt{(a_n^2+b_n^2)}\cos\left(\frac{2\pi n\,.\,t}{T_0}-\phi_n\right) \quad . \; . \; . \; . \; . \; . \; . \; . \quad (99)$$

where $\tan\phi_n=b_n/a_n$, and the spectrum may be plotted in terms of (1) the component amplitude $\sqrt{(a_n^2+b_n^2)}$ and (2) the phase angle $\phi_n$. Secondly, the complex exponential notation may be used as was explained in Sec. 10; we may write:

$$\left.\begin{aligned} a_n \cos\frac{2\pi n\,.\,t}{T_0} &= \frac{a_n}{2}(\epsilon^{j\omega_n t}+\epsilon^{-j\omega_n t}) \\ b_n \sin\frac{2\pi n\,.\,t}{T_0} &= \frac{b_n}{2j}(\epsilon^{j\omega_n t}-\epsilon^{-j\omega_n t}) \end{aligned}\right\} \quad . \; . \; . \quad (100)$$

where $\omega_n=2\pi n/T_0$.

This $n^{th}$ harmonic term, of frequency $\omega_n$, could be represented by the vector diagram of Fig. 13, where $\mathbf{C}=a_n/2$ and $\mathbf{S}=b_n/2$. The conjugate vectors in this case would be $\frac{1}{2}(a_n+jb_n)$ and $\frac{1}{2}(a_n-jb_n)$.

From equation 100 we may write this $n^{th}$ term as:

$$(a_n \cos \omega_n t + b_n \sin \omega_n t) = \left(\frac{a_n + jb_n}{2}\right)\epsilon^{-j\omega_n t} + \left(\frac{a_n - jb_n}{2}\right)\epsilon^{+j\omega_n t} \quad . \quad (101)$$

$$= \tfrac{1}{2}\sqrt{(a_n^2 + b_n^2)}\,[(\epsilon^{+j\phi_n})\epsilon^{-j\omega_n t} + (\epsilon^{-j\phi_n})\epsilon^{+j\omega_n t}] \quad . \quad (102)$$

where $\phi_n$ is given by equation 99.

However, we may write this most conveniently as the single term:

$$[\tfrac{1}{2}\sqrt{(a_n^2 + b_n^2)}\epsilon^{-j\phi_n}]\,\epsilon^{j\omega_n t} \quad . \quad . \quad . \quad . \quad (103)$$

if $n$ is now allowed to have *both positive and negative* integer values.

The whole series (96) for the complete wave may then be expressed as:

$$e = f(t) = \sum_{n=-\infty}^{n=+\infty} \alpha_n \epsilon^{j\omega_n t} \qquad \left(\text{where } \omega_n = \frac{2\pi n}{T_0}\right) \quad . \quad (104)$$

where $\alpha_n$ is a complex quantity. The D.C. component of the wave is still covered by the case of $n=0$.

Similarly the expression for $\alpha_n$, the complex amplitude, may be written:

$$\alpha_n = \tfrac{1}{2}\sqrt{(a_n^2 + b_n^2)}\,.\,\epsilon^{-j\phi_n} = \frac{a_n - jb_n}{2} = \frac{1}{T_0}\int_{-T_0/2}^{+T/2} f(t)\epsilon^{-j\omega_n t}.\,dt \quad . \quad (105)$$

from equation 97, which definies $a_n$ and $b_n$. Again $n$ has both positive and negative integer values. These two equations (104 and 105) should be compared with 96 and 97 using the trigonometrical notation.

We have so far dealt only with a periodic wave which has a known analytical form, but it is possible to determine the harmonics of a wave known graphically, though for which there is no given equation. Such analysis may be based on measured ordinates of the graph, spaced at known intervals, and is necessarily approximate, though may be made to any desired accuracy.

In these cases no exact integrals such as 97 or 105 may be formed, but some kind of graphical integration is necessary.[19, 21, 22] Many optical, electrical, and mechanical means for Fourier analysis have also been devised.[23, 24, 25]

## 21. Spectrum analysis of pulses and of single transients—the Fourier integral

It is frequently necessary to determine the form of the spectrum for a wave of *pulse type*—that is, a wave which has a finite value only

for a time short compared with its repetition period. (For example, see Fig. 28 (c).) If the repetition period becomes infinite the pulse becomes a single isolated transient (Fig. 28 (d)); such a transient must start at some definite instant of time, but may physically take an indefinite time to decay to zero, although in practice it may be considered to be substantially zero after a fixed length of time $T_1$—the pulse or transient *duration.*

The analysis of such repeated pulses, or single transients, follows logically from the Fourier series analysis of a periodic wave. In the case of a single transient, the "repetition period" being infinity, the fundamental frequency of the Fourier spectrum is zero. The harmonics are harmonics of "zero frequency," which means that the spectrum consists of terms which are indefinitely close together on a frequency scale, so that *a continuous spectrum* is obtained. There is a sinusoidal component at every conceivable frequency from zero to infinity, unless the particular shape of the transient requires certain gaps in the spectrum.

Since the energy of this single transient must be finite, it follows that in general all these spectrum components must be infinitesimally small, though they may have definite relative amplitudes.[19, 20]

In Fig. 27 there is shown a number of different types of pulse, all repeated at the interval $T_0$ and having a duration $T_1$. The ratio $T_0/T_1$ is written as $K$. Also the expression $e=f(t)$ for each pulse form is given, together with the expression $a_n$ for the spectrum components' amplitudes. Since all these pulses are symmetrical about $t=0$, cosine terms only are present, so that $b_n=0$ and $f(t)=f_c(t)$, since $f_s(t)=0$.

Taking the first waveform in this figure as an example—the half cosine wave—and writing the expression for this waveform, $e=f(t)=E \cos \pi t/T_1$ between limits $-T_1/2$ and $+T_1/2$ in the equation 97, to obtain the $n^{th}$ harmonic amplitude (of the fundametal frequency $1/T_0$), gives:

$$a_n=\frac{2}{T_0}\int_{-T_1/2}^{+T_1/2}\left(E\cos\frac{\pi\,.\,t}{T_1}\right)\cos\frac{2\pi n\,.\,t}{T_0}\,.\,dt\,. \quad . \quad . \quad (106)$$

Evaluation of this integral gives $a_n$, but we shall see that it is more simple to consider $a_nK$. The expression for $a_nK$ becomes, in this example:

$$a_nK=\frac{4E}{\pi}\cdot\frac{\cos n\pi/K}{[1-4n^2/K^2]}. \quad . \quad . \quad . \quad . \quad (107)$$

It may be seen that this expression is a function of $n/K$ and not of $n$ and $K$ separately. The same is true of the other examples listed in Fig. 27, as the reader may see, and indeed any pulse type of wave [26]

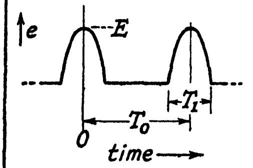
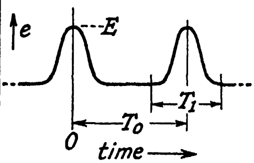
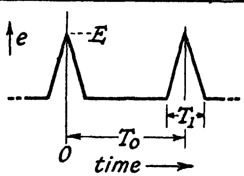
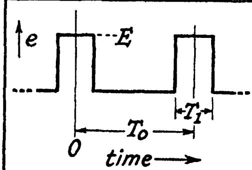
| Waveform | Wave Equation | nth Harmonic Amplitude |
|---|---|---|
|   (a) Half Cosine | $e = E\cos\frac{\pi}{T_1}t$ <br> $\left(-\frac{T_1}{2} < t < +\frac{T_1}{2}\right)$ | $a_nK = \dfrac{4E\cos\frac{n\pi}{K}}{\pi\left[1-\frac{4n^2}{K^2}\right]}$ |
|   (b) Cosine-squared | $e = E\cos\frac{2\pi}{T_1}t$ <br> $\left(-\frac{T_1}{2} < t < +\frac{T_1}{2}\right)$ | $a_nK = \frac{E.K}{\pi.n}\cdot\dfrac{\sin\frac{n\pi}{K}}{\left[1-\frac{n^2}{K^2}\right]}$ |
|   (c) Triangular | $e = E\left(1+\frac{2t}{T_1}\right)$ <br> for $-\frac{T_1}{2} < t < 0$ <br> $e = E\left(1-\frac{2t}{T_1}\right)$ <br> for $0 < t < +\frac{T_1}{2}$ | $a_nK = \frac{2E.K^2}{\pi^2.n^2}\left[1-\cos\frac{n\pi}{K}\right]$ |
|   (d) Rectangular | $e = E$ <br> $\left(-\frac{T_1}{2} < t < +\frac{T_1}{2}\right)$ | $a_nK = \frac{2E.K}{\pi.n}\left[\sin\frac{n\pi}{K}\right]$ |

Fig. 27.—Pulse Waveforms of Different Shapes, their Equations and Harmonic Amplitudes. [$K = T_0/T_1$; $T_0$=repetition period.]

having a finite repetition-period/duration ratio $T_0/T_1 = K$ may have its harmonic amplitudes expressed in the form:

$$a_nK = \text{function of } \left(\frac{n}{K}\right) \quad . \quad . \quad . \quad . \quad (108)$$

The spectrum of a pulse wave may therefore be plotted in terms of $n/K$ instead of $n$, so that it may be applied to a pulse of a given

waveform but of any ratio $K=T_0/T_1$. Fig. 28 shows in (a) a square wave, for which $K=2$, and its spectrum. This spectrum is plotted against $n$, the harmonic number with $1/T_0$ as the fundamental frequency, in terms of conjugates, and is identical with the spectrum in Fig. 25 (b). The dotted line on this spectrum is the envelope of the spectrum terms and is obtained by plotting the expression 92 (the square-wave spectrum) with $n$ as a continuous variable. The

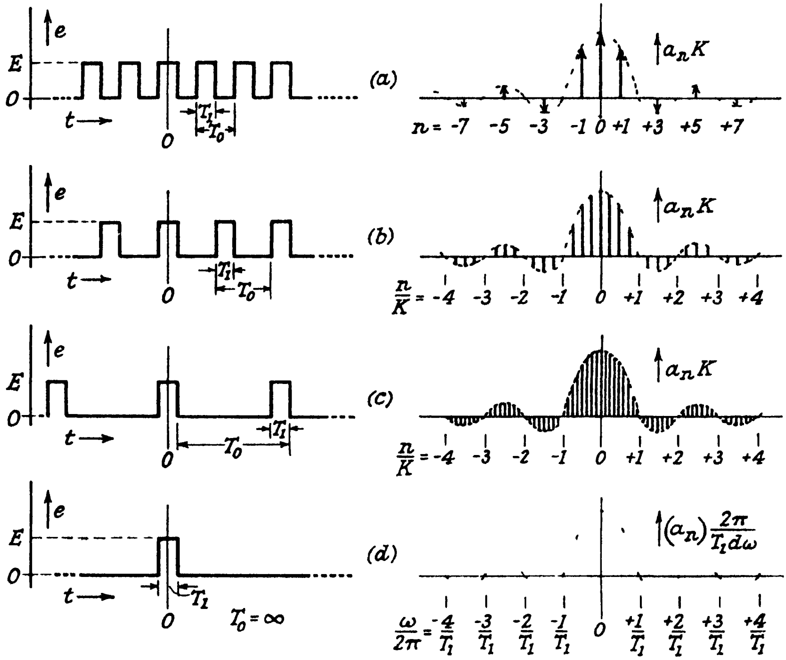



Fig 28.—Spectrum of a Rectangular Pulse, showing the Change as the Ratio $K(=T_0/T_1)$ is Increased The Series of Harmonic Terms becomes a Continuous Spectrum.

[The Spectra Curves are Diagrammatic only, and not Exact.]

actual harmonics in this case then occur at points where $n$ is an odd whole number. In Fig. 28 (b) the repetition period $T_0$ of the "square" wave has been doubled, so that it becomes a rectangular pulse wave of duration $T_1$, for which $K=4$. If the spectrum be plotted (as shown) against $n/K$, the same spectrum envelope applies:

$$a_nK=\frac{2EK}{\pi n}\sin\left(\frac{n\pi}{K}\right) \quad . \quad . \quad . \quad . \quad . \quad (109)$$

as may be seen by comparing this with the expression 92. This

expression 109 for the rectangular pulse wave may be derived by the reader in exactly the same way as the expression 107 was obtained; the value $f(t)=E$ over the limits $t=\pm T_1$ should be inserted in equation 97 and the integration carried out. The harmonics, however, will be spaced at half the interval in this case, compared with the square-wave case (a). Fig. 28 (c) shows the rectangular wave with a still greater value of $K$, and the spectrum plotted against $n/K$. The harmonics are much closer together, but still lie under the same spectrum envelope (109).

The vertical scale of these spectra is maintained by plotting $a_nK$ instead of $a_n$.

Thus a single "universal" spectrum envelope curve may be plotted for a given pulse shape independent of its repetition period. In the limit, $T_0 \to \infty$, a single pulse results, as in Fig. 28 (d), and the harmonics lie under this same spectrum envelope, but are continuous, being spaced at infinitesimal intervals.

As $T_0 \to \infty$ the harmonics come very close together, spaced by $2\pi/T_0$ on an angular frequency scale. If we call this $\delta\omega$, then $n\delta\omega$ becomes simply a continuous frequency scale $\omega$. Then

$$2\pi/T_0=\delta\omega, \quad K=T_0/T_1=\frac{2\pi}{T_1\delta\omega} \quad . \quad . \quad . \quad . \quad (110)$$

and

$$n=\frac{\omega}{\delta\omega} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (111)$$

so that the scales used in these "universal" spectra may be interpreted in this limiting case. Another way of looking at this horizontal scale $nT_1/T_0$ is that whereas $n$ is the harmonic number of the repetition period $T_0$, $nT_1/T_0$ is the harmonic number of the duration period $T_1$ even when $T_0 \to \infty$. The scales for the continuous spectrum will now be as given on Fig. 28 (d).

Now what happens to the Fourier series expressions, 96 and 97, in this case of an isolated pulse or transient? Substituting the continuous frequency variable $\omega$ for $2\pi n/T_0$ makes expression 96 represent a series of sinusoidal terms of amplitudes $a_n$ or $b_n$ whose frequencies are spaced at intervals ($\delta\omega$). The $\Sigma$ then becomes an integral. Equation 96 then reads:

$$f(t)=\int_0^\infty \left.\begin{matrix} a(\omega)\cos \\ b(\omega)\sin \end{matrix}\right\} \omega t \, . \, d\omega \quad . \quad . \quad . \quad . \quad (112)$$

Instead of discrete harmonic amplitudes, $a_n$ and $b_n$, these amplitude

functions have been written $a(\omega)$ and $b(\omega)$ to show they are continuous functions of $\omega$. Strictly $a_n/\delta\omega \to a(\omega)$ and $b_n/\delta\omega \to b(\omega)$.

These amplitude functions are obtained by making the same substitution in equation 97, remembering that these amplitudes are now all infinitesimal and that the integration limits $\pm T_0/2$ go to $\pm\infty$.

$$\left.\begin{matrix} a_n/\delta\omega \\ b_n/\delta\omega \end{matrix}\right\} \to \left.\begin{matrix} a(\omega) \\ b(\omega) \end{matrix}\right\} = \frac{1}{\pi}\int_{-\infty}^{+\infty} \left.\begin{matrix} f_c(t)\cos \\ f_s(t)\sin \end{matrix}\right\} \omega t \,.\, dt \quad . \quad . \quad . \quad (113)$$

The $\delta\omega$ shows that the amplitudes are infinitesimal. Here we have shown the waveform $f(t)$ divided into its even and odd components $f_c(t)$ and $f_s(t)$.

The two integrals 112 and 113, which correspond to the series expression 96 and integral 97, are called *Fourier transforms* and are of extreme importance in the understanding of network transient response.[27] It can be seen that they are almost symmetrical with respect to $\omega$ and $t$, but the symmetry is made more exact by adopting the conjugate component idea. Then both the integrals have limits $\pm\infty$. Splitting each cosine and sine term into two equal amplitude terms of equal positive and negative frequencies, as in 98, the integral 112 becomes:

$$\left.\begin{matrix} f_c(t) \\ f_s(t) \end{matrix}\right\} = \tfrac{1}{2}\int_{-\infty}^{+\infty} \left.\begin{matrix} a(\omega)\cos \\ b(\omega)\sin \end{matrix}\right\} \omega t \,.\, d\omega \quad . \quad . \quad . \quad (114)$$

which represents a continuous spectrum of components of all frequencies from $\omega = -\infty$ to $+\infty$. Such a spectrum is illustrated by Fig. 28 (d) for the symmetrical pulse wave shown.

The above integrals, 113 and 114, may be put in the exponential form if required by using equation 100, but although this form is advantageous for analytical work the above notation is really of more use for the geometrical aspect of the subject such as we are considering in this book.

In Fig. 29 (a) a number of different pulse waveforms are shown corresponding to those in Fig. 27 but with the repetition period $T_0 \to \infty$, all having a total duration $T_1$ and a maximum amplitude 1·0 units; in (b) the continuous spectra of these pulse waveforms are shown, plotted on a frequency scale in terms of $1/T_1$. The relative vertical scale is in terms of $a(\omega)K$, which is $a_n 2\pi/T_1\delta\omega$ in the limit as $T_0 \to \infty$, so that these same pulse and spectra diagrams may be used also for pulse waveforms repeated at regular time

intervals $T_0$ by putting $\delta\omega = 2\pi/T_1K = 2\pi/T_0$ (equation 110). In this case the spectra curves will represent the envelopes of the harmonic components, listed in the Fig. 27. Table A, below, lists the equations of these continuous spectra.

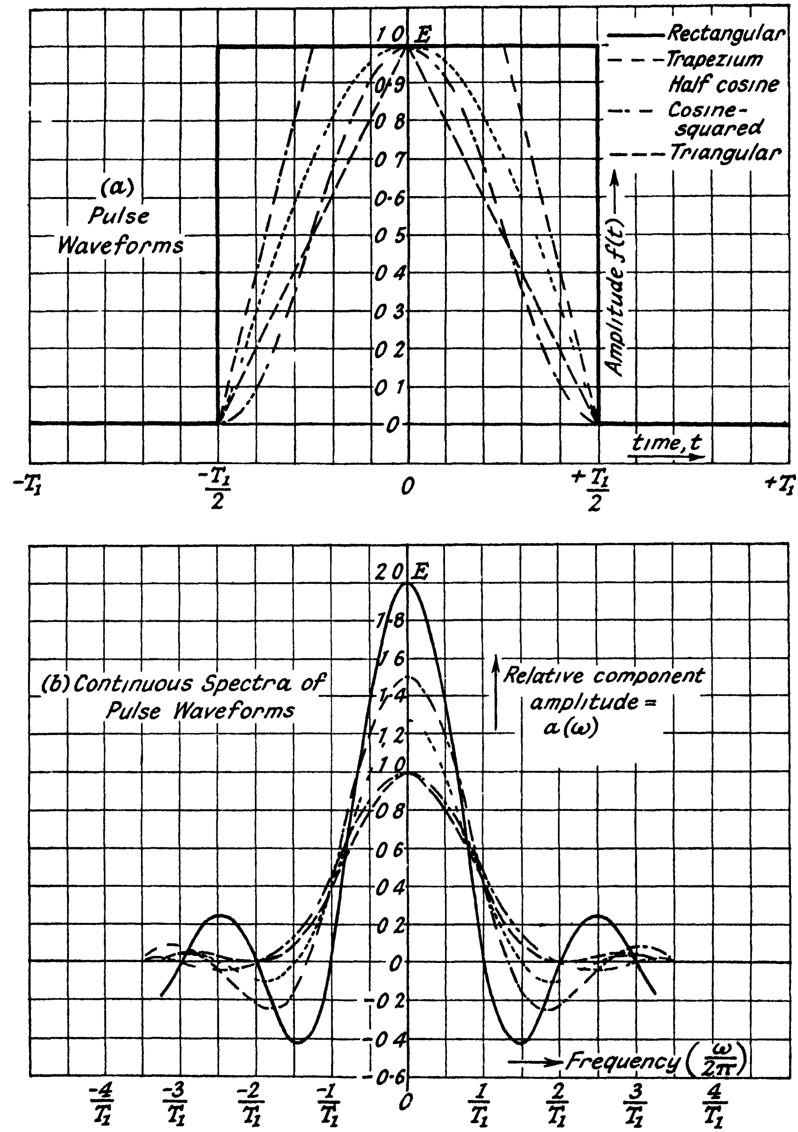



Fig 29.—Some Symmetrical Pulse Waveforms and their Spectra (cosine terms only)

*Table A.* (Fig. 29 (a) and (b))

| *Symmetrical Pulse Waveform* | *Spectrum Equation* |
|---|---|
| Rectangular . . | $a_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{4E}{\omega T_1}\left(\sin \frac{\omega T_1}{2}\right)$ |
| Half cosine . . . | $a_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{4E \cos \omega T_1/2}{\pi(1-\omega^2 T_1^2/\pi^2)}$ |
| Cosine-squared . . | $a_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{2E \sin \omega T_1/2}{\omega T_1(1-\omega^2 T_1^2/4\pi^2)}$ |
| Triangular . . . | $a_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{8E}{\omega^2 T_1^2}\left(1-\cos \frac{\omega T_1}{2}\right)$ |
| Trapezium . . . | $a_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{16E}{\omega^2 T_1^2}\left(\cos \frac{\omega T_1}{2} - \cos \frac{\omega T_1}{4}\right)$ |

*Table B.* (Fig. 29 (c) and (d), overleaf)

| *Skew-Symmetrical Waveform* | *Spectrum Equation* |
|---|---|
| Rectangular . . . | $b_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{-4E}{\omega T_1}\left(\cos \frac{\omega T_1}{2} - 1\right)$ |
| Triangular . . . . | $b_n \cdot \frac{2\pi}{T_1\delta\omega} = \frac{4E}{\omega T_1}\left(\frac{2 \sin \omega T_1/2}{\omega T_1} - 1\right)$ |
| Exponential . . . | $b_n \cdot \frac{2\pi}{T_1\delta\omega} = 4E\left(\frac{\omega T_1}{4+\omega^2 T_1^2}\right)$ |

The reader should study the shapes of these spectra curves in relation to the shapes of the corresponding pulse waveforms. The effects should be noted, on the spectra, of changing the slope of the pulse edge, of rounding the "corners" of the pulse waveform, etc.

Fig. 29 (c) shows some skew-symmetrical shapes of waveform, together with (d), their continuous spectra. Again the vertical scale is in terms of $b_n.2\pi/T_1\delta\omega$ and the diagrams may be used for the periodic forms of such waves. In this case the spectra are sine components since the waves are skew-symmetrical about the time origin. Table B, above, lists the equations for these continuous spectra.

Since the Fourier integral 113 is taken over an infinite period $-\infty < t < +\infty$, this method of calculating the frequency spectra $a(\omega)$ and $b(\omega)$ may be used for any transient waveform, which may not necessarily have a finite duration $T_1$ as is the case with the pulse-type waveforms. In practical circuits the transients that appear

must start at some definite time, but may also take an infinitely long time to cease, although in many cases the transient may be said to have died away to a negligible amplitude after a certain time—the *effective* duration of the transient. Any transient, what-

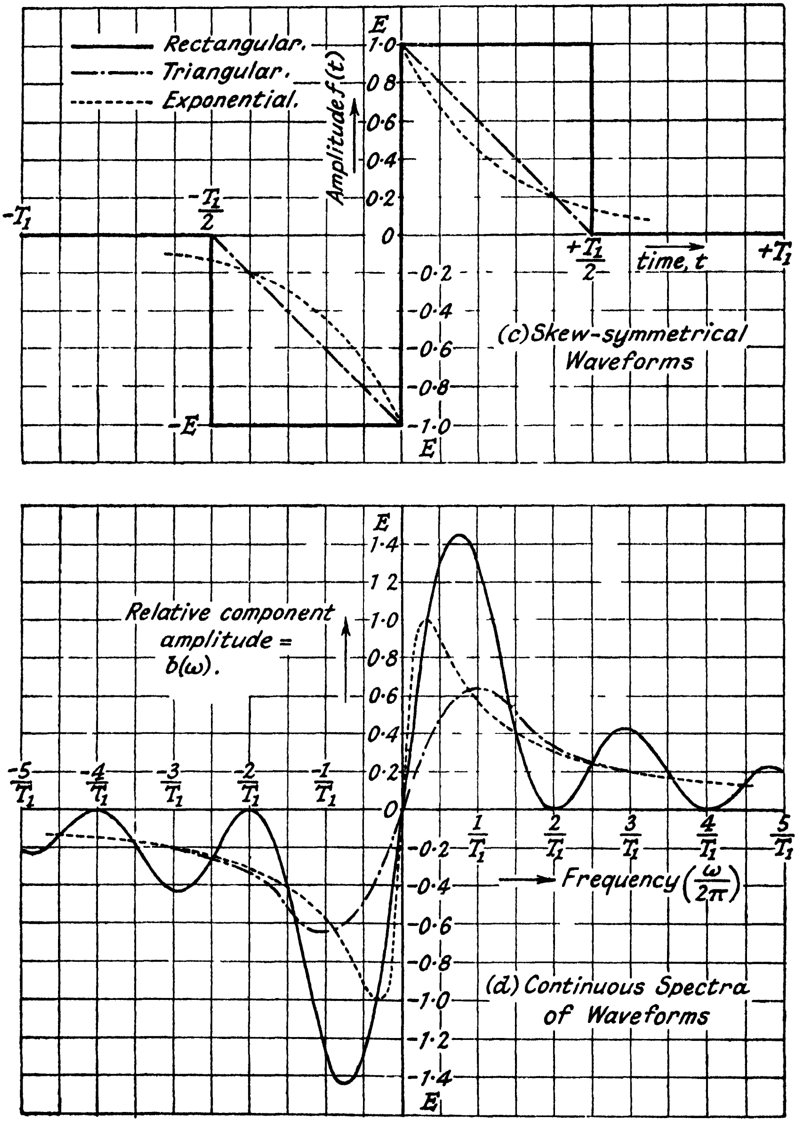



Fig. 29.—(c) Some Skew-symmetrical Waveforms and (d) Their Spectra.

ever its waveform, may be divided into odd and even components by applying the method discussed in Sec. 20 for periodic waves, and hence analysed into one spectrum of cosine components and another of sine components. Further examples of Fourier integral analysis of waveforms will be found in subsequent chapters, including two examples of particular interest, the rectangular pulse wave and the step wave, which are dealt with in Chapter 4, Sec. 35.

## 22. Some useful theorems concerning transient waveforms and their spectra

A few important points are mentioned in this section which will be dealt with in more detail, showing their practical applications, in later chapters.

(A) The theorem concerning the transfer of a carrier frequency to zero, which was discussed in Sec. 16, may be applied to pulse or transient envelope signals as well as to periodic waves. Thus any of the pulse spectra such as have been shown in Fig. 29 (b), plotted as conjugate components centred around a zero frequency (D.C.) component, could be transferred upon the frequency scale to be centred around some nominal carrier frequency $\omega_0/2\pi$. This transferred spectrum would then be the spectrum of a carrier wave of this frequency, modulated in amplitude by the original pulse. Figs. 18 and 19 illustrate the theorem as applied to a periodic envelope wave.

(B) The Superposition Theorem may be applied to spectra. Thus, just as we may add any number of different waveforms together to obtain a composite signal waveform, so may we add their spectra together, keeping cosine and sine components separate. The truth of this statement is too obvious to need proof. The time origins of the various waveforms added must not be changed of course; these origins must be as fixed for the calculation of the individual spectra. If $f_1(t)$, $f_2(t)$, $f_3(t)$ . . ., etc., are a number of different waveforms and $a_1(\omega)$, $a_2(\omega)$, $a_3(\omega)$ . . . their cosine component spectra and $b_1(\omega)$, $b_2(\omega)$, $b_3(\omega)$ . . . their sine component spectra, then the corresponding spectra of the composite waveform $f_1(t)+f_2(t)+f_3(t)$ . . . are given by: $a_1(\omega)+a_2(\omega)+a_3(\omega)$ . . . and $b_1(\omega)+b_2(\omega)+b_3(\omega)$ . . ., etc.

(C) In the last section it was pointed out that the two Fourier integrals, 113 (giving the spectrum envelope amplitudes $a(\omega)$ and $b(\omega)$) and 114 (giving the even $f_c(t)$ and odd $f_s(t)$ components of the transient waveform), were symmetrical with respect to time ($t$) and

angular frequency ($\omega$); that is to say, $t$ and $\omega$ could be interchanged without modifying the general forms of these integrals. This theorem has an important practical significance.

For example, in Fig. 28 (d) the rectangular-shaped pulse on the left has the spectrum envelope shown on the right; now according to the theorem the scales of $t$ and $\omega$ may be interchanged, and then a pulse of the shape shown on the right of this figure would have a spectrum envelope of the shape shown on the left (i.e. rectangular). The origins would remain unchanged. In general, if a waveform $f_c(t)$ has a spectrum envelope $a(\omega)$ *then a waveform* $a(t)$ *has the spectrum envelope* $f_c(\omega)$. Similarly for $f_s(t)$ and $b(\omega)$.

This interchange of a waveform and its spectrum* is one *of shape only*, and the relative amplitudes of the original waveform and the reciprocal one are not related, and indeed any arbitrary amplitudes may be chosen.

Again, any of the pulse waveforms in Fig. 29 (a) and the corresponding spectra envelopes (b) may be interchanged by writing $1/T_1$ for $T_1$, so that the curves in (b) would represent pulses and the corresponding curves in (a) would represent their spectra envelopes. Thus these diagrams serve a dual purpose.

For this reason the integrals 113 and 114 are referred to as Fourier *transforms*.

(D) Rayleigh's theorem states that the energy of a wave is the same whether it be calculated from its waveform or from its spectrum. Thus the energy $W$ of a wave having the waveform $f(t)$ is given by the expression:

$$W=\int_{-\infty}^{+\infty}[f(t)]^2 dt \quad . \quad . \quad . \quad . \quad . \quad . \quad (115)$$

where $W$ has suitable units.

Now the energy of a single sinusoidal component in the spectrum, of frequency $\omega/2\pi$ and amplitude $a(\omega)$ or $b(\omega)$, is proportional to $[a(\omega)]^2$ or $[b(\omega)]^2$. The total energy in the spectrum is given by the sum of the energy of every component. Therefore:

$$W=\int_{-\infty}^{+\infty}\begin{bmatrix}a(\omega)\\b(\omega)\end{bmatrix}^2 . \, d\omega \quad . \quad . \quad . \quad . \quad . \quad . \quad (116)$$

Since the integrals 115 and 116 are equal, this gives a relation

* Strictly speaking, it is *frequency*, $\omega/2\pi$ (not $\omega$), and *time*, $t$, that may be interchanged, since these are reciprocal quantities.

between the area under a waveform and under its corresponding spectrum envelope, with the ordinates squared in each case.

(E) In the case of a symmetrical pulse wave (such as in Fig. 29 (a)), the Fourier components being entirely cosine waves, the peak amplitude of the pulse at the time origin $t=0$ is equal to the sum of all the components in the spectrum, since $\cos 0=1$. In the case of a non-symmetrical pulse or transient, the peak height at $t=0$ is still equal to the sum of the cosine component amplitudes, since the sum of the sine components at $t=0$ is zero. Hence:

$$E=\int_{-\infty}^{+\infty} a(\omega)\,.\,d\omega \quad . \quad . \quad . \quad . \quad . \quad (117)$$

and this integral is equal to the area under the continuous spectrum envelope. Similarly the amplitude of a periodic wave, at the chosen origin $t=0$, is given by the sum of the (finite amplitude) cosine components in the harmonic spectrum.

## REFERENCES

1. F. M. Colebrook, "Alternating Currents and Transients," McGraw-Hill, 1912.
2. E. Guillemin, "Communication Networks," Vol. 1, John Wiley, New York (Chapman & Hall, London), Chapter III.
3. A. Bloch, "Modulation Theory," *J.I.E.E.*, Vol. 91, Part III, No. 13, March 1944, p. 31.
4. A. Russell, "Treatise on the Theory of Alternating Currents," Vol. 1, C.U.P., 1914, p. 419.
5. H. S. Carslaw, "Theory of Fourier's Series and Integrals," Macmillan & Co., London, 1930, Chapter 7.
6. R. V. L. Hartley, "Relations of Carrier and Sidebands in Radio Transmission," *B.S.T.J.*, Vol. 2, April 1923, p. 90.
7. T. E. Shea, "Transmission Networks and Wave Filters," D. Van Nostrand Co., New York, 1929, Chapter 10.
8. D. A. Bell, "F.M. Communication Systems," *Wireless Engineer*, May 1943, Vol. 20, p. 233. (An excellent brief summary of the main points of importance.)
9. H. Roder, "Amplitude, Phase, and Frequency Modulation," *Proc.I.R.E.*, Vol. 19, No. 12, Dec. 1931.
10. M. G. Crosby, "Communication by Phase Modulation," *Proc.I.R.E.*, Vol. 27, Feb. 1939, p. 126.
11. D. I. Lawson, "Frequency Modulation," *Wireless Engineer*, Vol. 17, Sept. 1940, p. 388.
12. M. G. Crosby, "Carrier and Side Frequency Relations, with Multi-tone Frequency or Phase Modulation," *R.C.A. Review*, Vol. 3, July 1938, p. 103.

13. E. C. CHERRY and R. S. RIVLIN, "Non-linear Distortion, with Particular Reference to the Theory of Frequency Modulated Waves," *Phil Mag.*, Vol. 32 (1941), p. 265. (Part I only—the general spectrum.)
14. N. W. MCLACHLAN, "Bessel Functions for Engineers" [see listed formulae, numbers 51–55], O.U.P., 1934.
15. VAN DER POL, "Frequency Modulation," *Proc.I.R.E.*, Vol. 19, 1930, p. 1194 (for spectrum of F.M. wave with square-wave modulation).
16. G. W. O. HOWE, *Wireless Engineer*, Editorials, Nov. 1939, Aug. 1940, Jan. 1941.
17. K. R. STURLEY, "Frequency Modulation," The Hulton Press, London, 1942.
18. A. HUND, "Frequency Modulation," McGraw-Hill, New York, 1942.
19. A. EAGLE, "Fourier's Theorem and Harmonic Analysis," Longmans, Green & Co., London, 1925.
20. H. S. CARSLAW, "Theory of Fourier's Series and Integrals," Macmillan & Co., London, 1930, Chapter 7.
21. A. CLAYTON, "Harmonic Analysis by Selected Ordinates," *J.I.E.E.*, Vol. 59, 1921, p. 491.
22. B. WEDMORE, "A Graphical Method of Analysing Harmonic Curves," *Electrician*, 1895, Vol. 35, p. 512.
23. R. FÜRTH and R. W. PRINGLF, "A New Photo-electric Method of Fourier Analysis," *Phil. Mag.*, Vol. 35, No. 249, Oct. 1944.
24. T. B. RYMER and C. C. BUTLFR, "An Electrical Circuit for Harmonic Analysis and other Calculations," *Phil. Mag.*, Vol. 35, No. 248, Sept. 1944.
25. C. A. BEEVERS, "A Machine for the Rapid Summation of Fourier Series," *Proc. Phys. Soc.*, Vol. 51, 1939, p. 660.
26. F. R. STANSEL, "Some Analyses of Waveshapes used in Harmonic Producers," *B.S.T.J.*, Vol. 20, July 1941, p. 331.
27. E. GUILLEMIN, "Communication Networks," Vol. II, John Wiley, New York (Chapman & Hall, London), p. 468.
28. R. V. L. HARTLEY, "A More Symmetric Fourier Analysis applied to Transmission Problems," *Proc. I.R.E.*, March 1942, Vol. 30, No. 3.

Although not referred to in this chapter, the following works are also of interest:

(A) G. JOOS, "Theoretical Physics," Blackie & Sons, 1934 (Chapter II, "Theory of Vibrations and Waves").
(B) J. R. CARSON and T. C. FRY, "Variable Frequency Electric Circuit Theory, with Application to the Theory of Frequency Modulation," *B.S.T.J.*, Vol. 16, Oct. 1937, p. 513 (advanced theory).
(C) J. R. CARSON, "Electric Circuit Theory and the Operational Calculus," Part III (derivation of Fourier Integral from Series, etc.).
(D) R. J. PIERACCI, "A Frequency-Modulation Monitoring System," *Proc.I.R.E.*, Vol. 28, Aug. 1940, p. 374 (frequency-modulation spectra).
(E) K. R. STURLEY, "Frequency Modulation," *J.I.E.E.*, Vol. 92, Part III, No. 19, Sept. 1945, p. 197 (see long list of included references).
(F) See *Nature*, Dec. 22, 1945, p. 756. Max Born, R. Fürth and R. Pringle.
(G) B. VAN DER POL, "The Fundamental Principles of Frequency Modulation," *J.I.E.E.*, Vol. 93, Part III, No. 23, May 1946, p. 153.

## CHAPTER 3

# THE STEADY-STATE CHARACTERISTICS OF NETWORKS

### 23. The measurement and specification of a network's characteristics

In practice all electric waveforms must be produced by some kind of electric circuit, and it will be well to consider the steady-state characteristics of circuits in general before passing on to their response to transients, pulses, and other complex waves.

Networks may be classified into various groups: thus they may be active or passive, linear or non-linear, and also they fall into various classes according to their frequency selective properties. Many of the networks used in communication channels consist of arrangements of reactances designed to have certain predetermined steady-state characteristics for the purpose of selecting signals which have frequencies lying between certain limits. Thus they may be used for separating signals of different waveforms; it is important to realise that if a network is able to discriminate between two coincident signals it can only do so on the basis of its steady-state characteristics. Such discrimination can never be exact and perfect, but the selective networks may attenuate the Fourier components of one signal more than those of another (unless these Fourier signals have coincident frequencies due to the signal's spectra overlapping). Such selective networks may also be used for the purpose of modifying the waveform of a transient in some desired fashion ("shaping" a wave, as it is sometimes called). One way of looking at this function of selective networks is from the steady-state point of view; the network modifies the relative amplitudes and phases of the Fourier components in the transient and thus changes the waveform. This is especially convenient in many cases, where the selective network design is based on certain standardised forms (as is frequently the case in communications) called *wave-filters*, which are classified and in common use. Their ideal steady-state characteristics, or "selectivity curves," are readily available, as are also the characteristics of other standardised networks which are not strictly to be called wave-filters, but which are frequently used.

The design of a wave-filter is based on the idea of a recurrent network of pure reactances, an example of which is shown in Fig. 30, which consists of an infinite number of identical sections. Each section is a *4-terminal* network, that is to say a signal applied to one pair of terminals, *ab*, is received at a second pair, *cd*.

If the *series-arm* and the *shunt-arm* of a filter section are dual impedances* the filter type is called a *constant-K* filter; this is clearly the case in the present example, the series-arm being an inductance and the shunt arm a capacity, which are duals. Other filter types are based on this fundamental constant-K type of structure, but we are not concerned with wave-filter design in this book and the reader is referred to existing literature on the subject.[1, 2, 10]

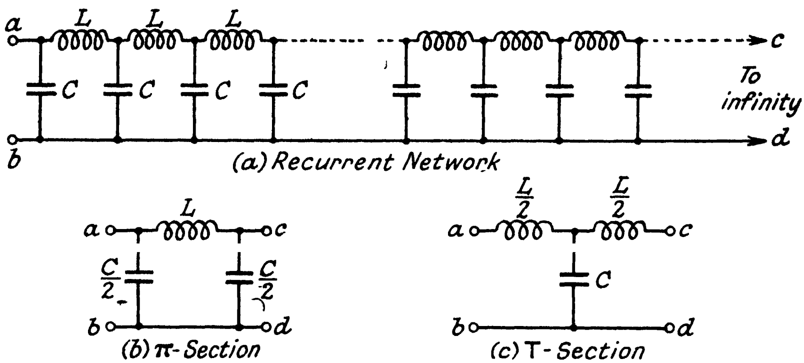



Fig. 30.—Filter Structures. A Recurrent Network may be Divided into T- or π-sections, considered to be Connected in Cascade.

Wave-filter types are classified according to their selectivity properties. Thus a filter which selects all signal components having frequencies less than a certain value is called a "low-pass filter," and is normally used for selecting envelope waves (as opposed to modulated waves). A filter which selects a continuous band of frequencies, thus being suitable for selecting amplitude-, phase-, or frequency-modulated waves, and which rejects frequencies outside this band, is called a "band-pass filter." Similarly those filters which select all frequencies above some fixed value or filters which reject all frequencies lying in a certain band are called "high-pass" and "band-elimination" filters respectively.

Wave-filters were originally designed for steady-state applications,

* See Sec. 2, Chapter 1.

as a means of selecting required signals in telephony systems. Recently these forms of network have come into use for communication systems in which transients are involved—telegraphy, television,[6] and radar—and calculations have been made, and published of the transient responses of some of these structures.[7, 8]

The published frequency characteristics of filters are usually idealised in that they refer to *infinitely long chains* of identical filter sections, as in Fig. 30 (a), whereas in practice only a few sections or, quite commonly, one only are used, together with a resistive termination.[3, 4, 5, 6] Fig. 31 shows some examples, (c) being two sections of a "constant-K" low-pass filter with a resistive termination, while (d) shows a single-section filter used as the anode load

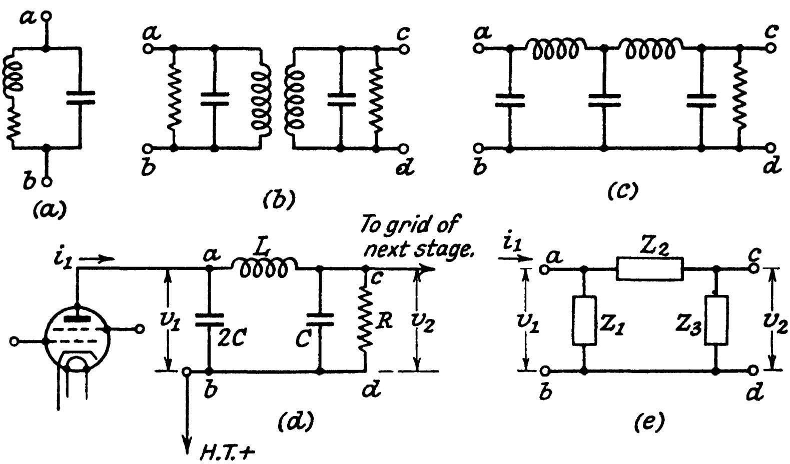



Fig. 31.—Some Simple 2- and 4-Terminal Networks.

of an amplifier stage. In such cases the frequency characteristics are only approximated by the idealised characteristics, and as regards an estimation of the transient response of such finite networks, their use can be very misleading. Also, the effects of dissipation in the elements is always neglected in the idealisation. In practice it is simplest either to calculate the steady-state response of these finite networks by use of Kirchhoff's laws or to construct the network and make a measurement of the characteristics.

In the same Fig. 31, (a) shows a single-tuned circuit and (b) a typical band-pass filter as used in television amplifiers. All these networks in this figure are what we have called "standardised structures." (e) represents a schematic 4-terminal network, built

up as a $\pi$-section, of three 2-terminal impedances $Z_1$, $Z_2$, and $Z_3$, which may take on any desired physical form.

In order to specify the behaviour of any particular network, as regards its selective properties or its response to an E.M.F. of given waveform, it is insufficient to say that it belongs to a certain class of wave-filter or other standardised network. If the network be a simple series tuned circuit, then in such a case it is sufficient to say that it is (for example) a resonant circuit of a certain Q; this determines the waveform of its response to any applied E.M.F., or injected current, of known waveform. But if the network is a more complex type, possessing a number of meshes, then a knowledge of its particular class or form may give some general indication of its response, but only in a qualitative way.

Exact determination of a network's response to a given signal of known waveform, transient or periodic, is possible if the steady-state response of the network has been measured at all frequencies from zero to infinity. In practice it is usually necessary only to measure the response at certain frequency intervals and then to interpolate the response at the intermediate frequencies; also there is always some limit to the frequency range that needs to be covered.

Sometimes, for example in telephony, it is customary to measure only the *magnitude* of the network's steady-state response to signals of given frequencies, since the phase-shift is not of importance.* In other cases, both characteristics must be measured—magnitude (sometimes called amplitude or modulus) and phase-shift [9, 11]; for example, this is most important in television channels, through which transients are passed possessing definite waveforms which must be preserved within narrow limits. The response of a channel to a wave having a specified waveform may be calculated provided that *both parts* of the steady-state characteristics are known, the modulus and the phase-shift, at all frequencies. In practice there are alternative methods of assessing the phase-shift distortion of a channel, which are useful since it is extremely difficult to measure phase-shift directly. These methods depend upon observing the waveform distortion of some transient wave of definite shape, for example a step wave as in Fig. 4 (a).

* The ear, unlike the eye, cannot detect relative phase-shift between pure tones, but responds only to their relative amplitudes. Thus for electrical transmission of sound signals the phase-shifts through the network employed are not, within reasonable limits, important.

## 24. Definitions of the characteristics of two- and four-terminal networks

Many networks used in communication channels are 4-terminal, in that a wave is applied to one pair of terminals and received at a second pair. If the network contains reactive elements the received waveform on the output terminals will, in general, differ in shape from the applied waveform on the input terminals. Such a network may be composed of a number of branches each of which will be a 2-terminal impedance. Fig. 31 illustrates some examples: (a) is a 2-terminal impedance, whilst (b) and (c) are 4-terminal.

As regards a specification of the characteristics of a 4-terminal

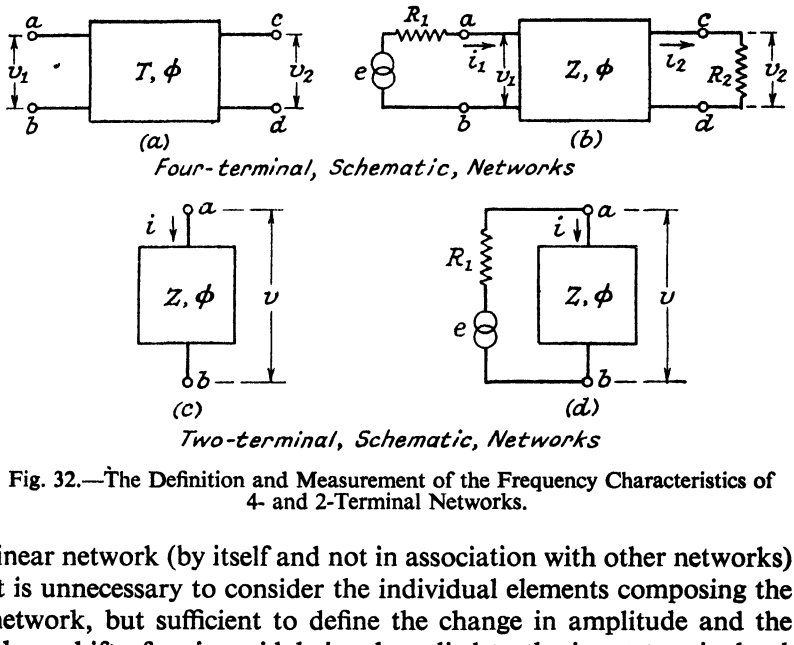



Fig. 32.—The Definition and Measurement of the Frequency Characteristics of 4- and 2-Terminal Networks.

linear network (by itself and not in association with other networks) it is unnecessary to consider the individual elements composing the network, but sufficient to define the change in amplitude and the phase-shift of a sinusoidal signal applied to the input terminals *ab* and received on the output terminals *cd*. Thus it is convenient to represent the network by a schematic "box," as in Fig. 32 (a), which may contain any number and arrangement of elements. Obviously, in order to be able to *calculate* the network's characteristics, one must know the internal structure of the box.

The 4-terminal network (a) has a voltage $v_1$ across its input terminals *ab* and a voltage $v_2$ across its output terminals *cd*. The network is assumed to be linear, that is to say, the amplitude of $v_2$

is proportional to the amplitude of $v_1$, and the phase difference between $v_2$ and $v_1$ is independent of these amplitudes, at any particular frequency. If $v_1$ is sinusoidal (as we assume here), then $v_2$ must also be sinusoidal.

In general, $v_2$ will differ from $v_1$ both in amplitude and phase, depending on the arrangement of elements inside the box.

$$\text{Let} \left.\begin{aligned} v_1 &= V_1 \cos \omega t \\ v_2 &= V_2 \cos (\omega t - \phi) \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (118)$$

Then $v_2$ (as written) lags behind $v_1$ by an angle $\phi$ or a time $t_1$ ($=\phi/\omega$ since $v_1$ rises to its crest value at $t=0$ and $v_2$ later at $t=+\phi/\omega$). Thus we may consider $v_2$ to have a negative phase-shift ($-\phi$) relative to $v_1$.

If we write $V_2/V_1=T$ (the ratio of the amplitudes of $v_2$ and $v_1$), then the two quantities $T$, $-\phi$, may be called the *transfer constants* at the particular frequency $\omega$, in this case measured in terms of $v_2$ relative to $v_1$. Thus we may consider ($T$, $-\phi$) as the ratio $v_2/v_1$ or output/input transfer ratio.

However, it is the usual practice in telephony, in line telegraphy, and wherever transmission lines are used, to define the transfer constant of a 4-terminal system by the inverse ratio $v_1/v_2$ or input/output. In the present example such a definition would give:

$$\frac{V_1}{V_2}=\frac{|v_1|}{|v_2|}=\frac{1}{T} \quad . \quad . \quad . \quad . \quad . \quad . \quad (119)$$

Phase of $v_1$ relative to $v_2=+\phi$.

It is not of great importance which definition is taken, but whenever the characteristics of a network are written down, care should be taken to state whether they represent $v_2/v_1$ or $v_1/v_2$.

The definitions above have been made for one fixed frequency. If now $T$ and $\phi$ be measured at all frequencies they will both (in general) be functions of frequency and could be written $T(\omega)$ and $\phi(\omega)$, and would then represent the frequency characteristics of the network in the particular form $v_2/v_1$. It may appear wrong to call $T(\omega)$ and $\phi(\omega)$ transfer *constants*, since they are functions of frequency, but it is the usual name; sometimes, for this reason, they are called transfer functions.

For analytical work it is often advantageous to use the exponential notation; for example, we may write equation 118 as:

$$\left.\begin{aligned} v_1 &= \text{real part of } V_1\epsilon^{j\omega t} \\ v_2 &= \text{real part of } V_2\epsilon^{j(\omega t-\phi)} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (120)$$

Then the transfer function $v_2/v_1$ = real part of $(V_2/V_1)\epsilon^{-j\phi}$ and since in general $V_2/V_1$ and $\phi$ vary with frequency this quantity is usually written as a complex function of $\omega$, thus:

$$\text{Transfer function} = F(j\omega) \quad . \quad . \quad . \quad (121)$$

Fig. 32 (b) shows the 4-terminal network connected up to a generator (sinusoidal wave) of internal impedance $R_1$ and E.M.F. $e$, and with a termination $R_2$ which is resistive in this instance. The currents $i_1$ and $i_2$ flow in series with the input and output terminals of the network respectively. Then all the quantities $i_1$, $i_2$, $v_1$, $v_2$ may be measured and used to define the transfer-function characteristics of the network, in a number of ways. Thus we may specify:

(1) $v_2/v_1$—as we have already discussed,
(2) $v_2/i_1$,
(3) $i_2/i_1$,

at any or every frequency, as an analytical function or a measured set of values, and similarly for the three inverse characteristics.

It may be seen that the form (2) above has the units of impedance, being the ratio of the output voltage across terminals *cd* to the input current in the external loop joining *ab*. This particular characteristic form, whether it be a calculated function of $\omega$ or a practical measured set of curves, is called the *transfer impedance*. Its inverse form, $i_1/v_2$, is similarly called the *transfer admittance*.

In the diagram, Fig. 32 (b), the network in the "box" is shown as having a transfer impedance of magnitude $Z$, $\phi$

$$Z = \frac{|v_2|}{|i_1|} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (122)$$

and phase angle $\phi$, being the phase of $v_2$ relative to $i_1$.

The choice of the form of characteristic to be measured or calculated for a network, in any particular case, depends on the use to which the information is to be put. We may perhaps wish to determine the output signal from the network corresponding to a signal applied across the input terminals *ab* when this signal is given either as a current or as a voltage of known waveform. Again, the output signal may need to be known either as a current or as a voltage.

The relation between the input current $i_1$ and voltage $v_1$ (as the ratio $v_1/i_1$) is the *input impedance* of the network, being the impedance measured between the terminals *ab*. Similarly the ratio $v_2/i_2$ is the *output impedance*. The three impedances input, output,

and transfer, if known both as regards amplitude and phase-shift, are sufficient to define the properties of a 4-terminal network when it is used in conjunction with other networks. They determine the steady-state response and hence, if known for all frequencies from zero to infinity, determine also the transient response of the network.

Turning now to a 2-terminal network, Fig. 32 (c) illustrates such a unit in "box" form, which represents any arrangement of elements having any number of meshes, but only two accessible terminals. Fig. 32 (d) shows the network connected up to a generator of internal impedance $R_1$ and E.M.F. $e$. The steady-state behaviour of such a structure may be specified by two forms of characteristics, either by the ratio of the terminal voltage to the input current $i$, as $v/i$, or by the inverse ratio $i/v$. The former is, of course, the direct impedance and the latter the direct admittance of the network, measured between the terminals *ab*. The impedance "looking into" a network (that is, the impedance measured across the input terminals of the 2- or 4-terminal networks in the diagrams) is sometimes called the *driving-point impedance* since it is the impedance which the network presents to the generator.

$$\text{If } \left.\begin{aligned} i &= I\cos\omega t \\ v &= V\cos(\omega t-\phi) \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (123)$$

the impedance "frequency characteristics" of this 2-terminal network may be defined by the magnitude $V/I$ and by the phase angle of $v$ relative to $i$, that is $-\phi$.

## 25. Conjugate component characteristics

As an example of plotted frequency characteristics of a network, Fig. 33 (c) shows the transfer impedance characteristics $v_2/i_1$ of the network illustrated by Fig. 31 (d).

These characteristics may be calculated by a conventional method, as follows:

The $\pi$ block diagram network of Fig. 31 (e) represents the filter section (d) schematically, where:

$$\left.\begin{aligned} Z_1 &= \frac{1}{j.2\omega C} \\ Z_2 &= j\omega L \\ Z_3 &= \frac{R}{1+j\omega CR} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (124)$$

Now $$v_2/v_1 = \frac{Z_3}{Z_2+Z_3}$$

and $$v_1 = i_l\left[\frac{Z_1(Z_2+Z_3)}{Z_1+Z_2+Z_3}\right]$$

Hence $$v_2/i_1 = \frac{Z_1Z_3}{Z_1+Z_2+Z_3} \quad . \quad . \quad . \quad . \quad . \quad (125)$$

$=$the transfer impedance of the network.

Substituting the component values from 124 and simplifying,

$$\frac{1}{R}\cdot\frac{v_2}{i_1} = \frac{(1-2\Omega^2k)-j(3\Omega-2\Omega^3k)}{(1-2\Omega^2k)^2+(3\Omega-2\Omega^3k)^2} \quad . \quad . \quad (126)$$

where $\Omega = \omega CR$ and $k = L/CR^2$.

The expression 126 gives the real and imaginary parts of the transfer impedance; the amplitude and phase-shift components, as an alternative, may be found from these. Writing these real and imaginary parts as $C$ and $jS$ respectively:

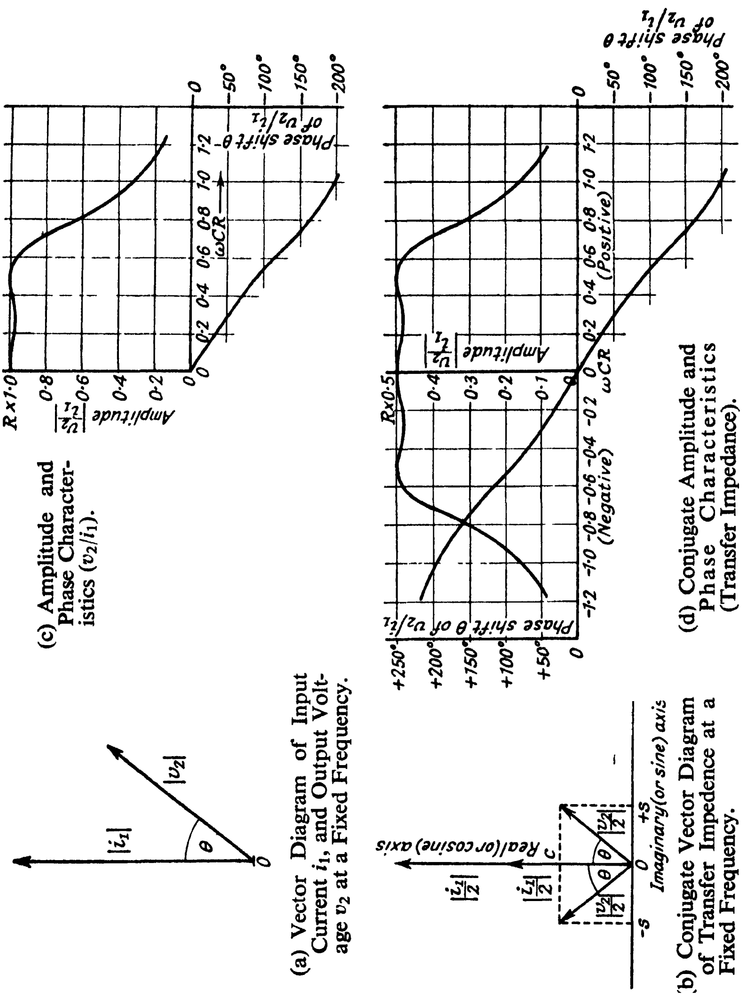
$$\frac{1}{R}\cdot\frac{v_2}{i_1} = (C+jS)$$

Then $$\left.\begin{aligned}\frac{1}{R}\cdot\frac{v_2}{i_1} &= \sqrt{(C^2+{}^2S)}\\ \text{and}\qquad \tan\theta &= S/C\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (127)$$

The constant $k$ in equation 126, has the units of $Q^2$ where Q is the factor associated with a simple series tuned circuit, Fig. 1 (a), for defining its energy storage:dissipation ratio.* The network here is not such a simple circuit, and strictly speaking Q has no simple meaning, but it is a very convenient constant in many cases of such filter sections or other similar networks.[15] Thus by giving $k$ different numerical values, the characteristics of the general network structure may be calculated with various relative values between the elements. Similarly the variable $\Omega$, proportional to frequency, may be used as a "relative frequency" so that one set of characteristics may be plotted for the structure, such that the curves are relevant to that particular *structure* and are independent of the relative values of $C$ and $R$. In Fig. 33 (c) the amplitude and phase-shift characteristics (equation 127) are plotted from zero frequency up to well above the "cut-off" frequency of this filter section. Such generalised curves are known as ***universal characteristics.*** In the set plotted in Fig. 32 (c) the value of $k = L/CR^2$ is 2.

* See Sec. 5 (c), Chapter 1, and equation 39.

The vector diagram showing $v_2$ and $i_1$ may be drawn for any particular frequency. Fig 33 (a) shows the vector $v_2$ lagging by the angle $\theta$ relative to $i_1$. Now the vector representing the injected current $i_1$ is the datum vector and may be regarded as being of the




(a) Vector Diagram of Input Current $i_1$, and Output Voltage $v_2$ at a Fixed Frequency.

(b) Conjugate Vector Diagram of Transfer Impedence at a Fixed Frequency.

(c) Amplitude and Phase Characteristics ($v_2/i_1$).

(d) Conjugate Amplitude and Phase Characteristics (Transfer Impedance).

Fig. 33.—Vector Diagrams and Frequency Characteristics of the Transfer Impedance of a 4-Terminal Network.

[(a) and (c) Single components; (b) and (d) Conjugate components.]

fixed length shown, $I_1 = |i_1|$, and in a fixed position at all frequencies. At different frequencies the vector $v_2$ then has different magnitudes and phase angles relative to $i_1$; if its length be made equal to $|v_2/i_1|$ the locus of this vector, as the frequency is varied, would represent

the transfer impedance characteristics of Fig. 33 (c) plotted on polar coordinates.

At any particular frequency the vector $v_2/i_1$ may be drawn in terms of two conjugate components, in the same way that the vector for a sinusoidal wave was represented by two conjugate components in Fig. 13.

Dividing the vector $v_2$ in half and plotting two vectors of magnitude $|v_2/2|$ at an angle $\theta$ on either side of the real axis gives, for one particular frequency, the conjugate vector diagram of Fig. 33 (b). The input current vector $i_1$ lies along this real axis and its two conjugate components add up arithmetically to the original length $|i_1|$. If these vectors $|v_2/2|$ now be divided by the constant magnitude $|i_2/2|$ they will represent the transfer impedance vectors as a conjugate pair.

Again the magnitudes and phase-angles of such conjugate vectors may be plotted on the linear frequency scale $\omega CR$ as in Fig. 33 (d); the right-hand vector (being a positive frequency vector) gives the positive frequency part of these frequency characteristics, while the left-hand vector gives the "negative frequency" part. As when using conjugate components for representing frequency spectra, there is no need here to worry what is meant, physically, by "negative frequencies"; these characteristics represent the transfer impedance of the network in the same way as do those in Fig. 33 (c), except that they are in a more symmetric form, the advantages of which will appear later. These conjugate characteristics may be obtained by plotting the magnitude of $v_2/i_1$, at half scale, and as a *mirror image*, about the origin or zero frequency point, and the phase-shift curve as a *skew mirror image* about this point.

Similarly, if the input current $i_1$ be of constant magnitude $I_1$ and considered to be at datum phase, it is given by

$$i_1 = I_1 \cos \omega t \qquad \qquad (128)$$

then the output voltage $v_2$ will be:

$$v_2 = V_2 \cos(\omega t - \theta) \qquad \qquad (129)$$

and this may be split into equal positive and negative frequency components:

$$V_2 \cos(\omega t - \theta) = \frac{V_2}{2} \cos(\omega t - \theta) + \frac{V_2}{2} \cos -(\omega t - \theta) \qquad (130)$$

The ratio of each of these conjugate voltage components to the input current $i_1$ gives the conjugate transfer impedance characteristics,

represented by Fig. 33 (d), having equal amplitudes $V_2/2I_1$ and phases $\pm\theta$ respectively.

## 26. Real and imaginary components of network characteristics

In a manner similar to that used in the preceding chapter for representing the spectra of waves, we may plot the characteristics of networks in two component parts—a symmetrical real component and a skew-symmetrical imaginary component. For instance, consider again the input current $i_1$ and output voltage $v_2$ of the 4-terminal network as given by equations 128 and 129. The voltage $v_2$ has been divided into two conjugate components in 130 and these are represented vectorially by Fig. 33 (b). Each of these vectors has two orthogonal components, since:

$$\frac{V_2}{2}\cos(\omega t-\theta)=\left[\frac{V_2}{2}\cos\theta\right]\cos\omega t+\left[\frac{V_2}{2}\sin\theta\right]\sin\omega t \quad . \quad (131)$$

which are represented in the diagram by $OC$ and $OS$ for the right-hand vector, with $OC$ and $-OS$ for the left-hand vector, where:

$$OC=\left[\frac{V_2}{2}\cos\theta\right], \qquad OS=\left[\frac{V_2}{2}\sin\theta\right] \quad . \quad . \quad (132)$$

The component $OC$ is the amplitude of a cosine wave and hence this is in phase with the applied current $i_1$ (equation 128), while the component $OS$ is orthogonal to this.

Calculating these components, given by 132, from the amplitude and phase-shift characteristics over the whole range of the frequency scale and dividing each by the constant magnitude $I_1(=|i_1|)$, we may plot the result as the real $[C]$ and imaginary $[jS]$ components of the transfer impedance of the network.

Fig. 33 (e) shows these components for our particular network example (Fig. 31 (d)), whose amplitude and phase-shift characteristics are plotted in Fig. 33 (d). Note that, when drawn in the conjugate form, the real component $[C]$ is symmetrical about zero frequency, while the imaginary component $[jS]$ is skew-symmetrical.

When using a set of frequency characteristics for calculating the distortion of a signal, the component parts of the characteristic may be taken separately. For example, the schematic "boxes" in Fig. 34 (b) represent the amplitude and phase-shift components of a 4-terminal network's characteristics, as separate units in cascade. The network's terminals are external, *ab*, *cd* as shown, and this division into two units is quite hypothetical, since no network can

be made which is selective as regards its amplitude characteristic but at the same time has a purely uniform phase characteristic (the converse is not true, since networks can be constructed which have non-uniform phase characteristics, but constant, non-selective, amplitude characteristics).*

For the purposes of calculation, we may apply the input signal to the input terminals *ab*, determine the effect of the amplitude characteristic $T$ alone, thus giving the hypothetical waveform on the "terminals" *ef*, and then add the effect of the phase-shift characteristic $\phi$, thus giving the true output of the network $v_2$ on

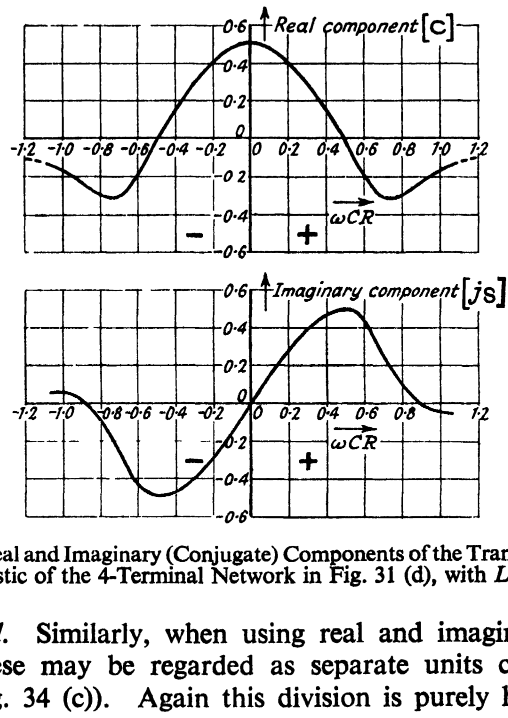



Fig. 33 (e).—Real and Imaginary (Conjugate) Components of the Transfer Impedance Characteristic of the 4-Terminal Network in Fig. 31 (d), with $L/CR^2=2{\cdot}0$.

terminals *cd*. Similarly, when using real and imaginary characteristics, these may be regarded as separate units connected in parallel (Fig. 34 (c)). Again this division is purely hypothetical, but is useful for calculation purposes. Thus we may calculate the response of the real component ($T\cos\phi$, which has zero phase-shift at all frequencies) to the applied wave and then add the response of the imaginary component ($jT\sin\phi$, which has $\pi/2$ phase-shift at all frequencies), giving the true network output signal $v_2$ on terminals *cd*. These conjugate characteristics of a low-pass filter, as drawn in Fig. 33 (d) (amplitude and phase-shift components) or

* See Sec. 39.

Fig. 33 (e) (real and imaginary components) may be regarded as the characteristics of a band-pass filter with *zero frequency midband*, since they are of symmetrical form about zero frequency. This ideȧ has considerable practical significance when we deal with the distortion of amplitude-modulated waves, and will be examined further in Sec. 41.

## 27. Four-terminal networks in cascade

A communication channel consists essentially of a number of 4-terminal network units connected in cascade. This is the simplest way of regarding a complete channel to be built up. Fig. 34 (a) illustrates such a chain of units, numbered 1 to $N$, which in practice may consist of band-pass filters, lengths of transmission line, amplifiers, or other types of network. Other forms of connection may be needed—for instance feedback may be provided, or a subsidiary chain of networks may branch off from the terminals of one of these units. But this is the basic, schematic, arrangement in which a signal is impressed on the input terminals *ab* of the first stage and transmitted through to the output terminals *cd* of the $N^{th}$ stage, undergoing distortion in each unit, depending upon the amplitude characteristic $T$ and phase-shift characteristic $\phi$ of each stage* (limiting our considerations to *linear* networks for the moment); these frequency characteristics may be different for each unit.

Now the characteristics (for example, of the first unit) may have been measured with the unit taken out of the complete channel and terminated by some nominal impedance, as in the diagram Fig. 32 (b) which shows a resistive termination. Such a termination may not be equal, in impedance, to the input impedance of the second unit with which the unit being measured will have to work in practice. These characteristics, as measured, will therefore not represent the behaviour of this unit in practice but, if they are to be correct, the characteristics must either be measured with a termination equal in impedance to the input impedance of the second unit or they must be measured with the first unit *in situ*. Similarly for the other units in the channel.†

* $T$ and $\phi$, functions of frequency, are again written short for $T(\omega)$ and $\phi(\omega)$.

† There are methods of calculating the change in characteristics due to mismatched terminations, with standard filter networks, by the use of what are called "reflection and interaction factors."[10]

In the chain of units shown in Fig. 34 (a) the transfer ratios $(T_1, \phi_1)$, $(T_2, \phi_2)$ . . ., etc., are assumed to apply to the units *in situ*, so that the transfer ratios $v_{n+1}/v_n$ have the magnitudes:

$$\left|\frac{v_{n+1}}{v_n}\right| = T_n \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (133)$$

And the phase-shift of $v_{n+1}$ relative to $v_n$ is

$$\psi_n = \phi_n \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (134)$$

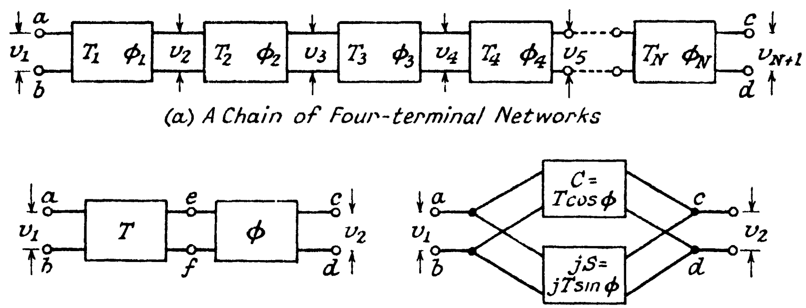
The transfer ratio of the whole chain, $v_{N+1}/v_1$, in terms of the individual ratios, thus has the magnitude:

$$\left|\frac{v_{N+1}}{v_1}\right| = \left|\frac{v_2}{v_1} \cdot \frac{v_3}{v_2} \cdot \frac{v_4}{v_3} \; . \; . \; . \; \frac{v_{N+1}}{v_N}\right| = (T_1 \, . \, T_2 \, . \, T_3 \; . \; . \; . \; T_N). \quad (135)$$

while the phase-shift of $\dot{v}_{N+1}$ relative to $v_1$ is clearly:

$$\phi_{TOTAL} = (\phi_1 + \phi_2 + \phi_3 + \; . \; . \; . \; \phi_N) \quad . \quad . \quad (136)$$

Again each unit in the chain may be considered to consist of



(a) A Chain of Four-terminal Networks

(b) Modulus and Phase "Components" (c) Real and Imaginary "Components"

Fig. 34.—Illustrating the Division of a Channel into a Number of "Components."

two hypothetical units in cascade (Fig. 34 (b)) so that the overall amplitude characteristic, equation 135, and phase-shift characteristic, 136, may be regarded as separate entities for the purpose of calculation of the overall network-chain response.

Similarly, each unit of the chain may be split into its real [$C$] and imaginary [$jS$] parts (Fig. 34 (c)), and the overall characteristics are then, $Z_N(j\omega)$:—

$$Z_N(j\omega) = [C_1 + jS_1]\,[C_2 + jS_2]\,[C_3 + jS_3] \; . \; . \; . \; . \; . \; . \quad . \quad . \quad (137)$$

which is not readily reduced to real and imaginary parts. However, if this be expressed in exponential form:

$$Z_N(j\omega)=T_1\,\epsilon^{j\phi_1}\,.\,T_2\,\epsilon^{j\phi_2}\,.\,T_3\,\epsilon^{j\phi_3}\,\ldots\ldots$$

then by taking logarithms:

$$\text{Log}\,Z_N\,(j\omega)=\Sigma\,\text{Log}\,[C_n+jS_n]=\Sigma\,\text{Log}\,T_n+j\Sigma\phi_n\ . \quad (138)$$

## 28. Equivalent networks and dual networks

Networks which have identical "frequency characteristics" are said to be *equivalent*, though this equivalence may be either *complete* or *limited.* Completely equivalent 4-terminal networks may be defined as having identical input, output, and transfer impedance* characteristics at all frequencies. Their external behaviours are then identical and any one may be substituted for any other in a complete system (e.g. the chain of networks in Fig. 34 (a)), and each has the same effect on the overall characteristics of the system. The differential equations of such networks, representing their transient behaviour, are identical.

The condition for complete equivalence of 2-terminal networks is that their driving-point impedances (modulus and phase-shift) should be equal at all frequencies. Since 4-terminal networks may be considered to be built up of a number of 2-terminal networks, their equivalence depends upon the possibility of constructing equivalent 2-terminal networks, but only to a limited extent.

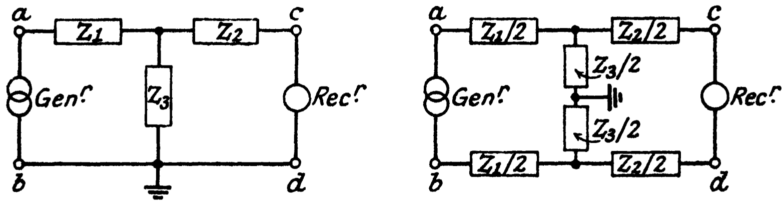
The reader should beware of the use of the term "equivalent network" in connection with circuits which are equivalent at *one frequency only.* For example, it is occasionally useful in continuous-wave work to use the damped tuned circuits in Fig. 45 as equivalents at the resonant frequency, but their characteristics are completely different at every other frequency. Obviously the transient responses of such networks must differ, since for them to be identical would necessitate equivalence at every frequency from zero to infinity (or at least covering the range of frequencies in the applied transient).

As an example of a *limited* equivalence between two networks we may quote the unbalanced and balanced forms of a given network. A balanced network is one in which the various impedance elements are connected in a symmetrical manner with respect to earth (or some "neutral" line). In an unbalanced network, there is a conducting, impedanceless connection between one input and one output terminal. Such networks, which are limited equivalents,

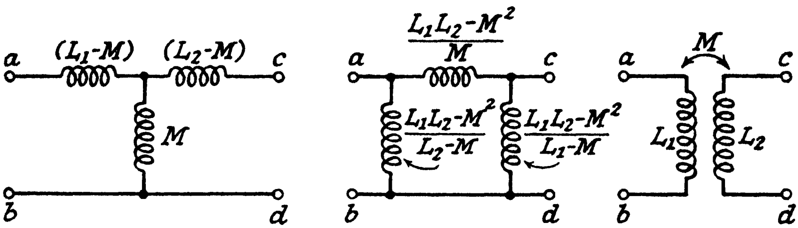
* Three network parameters must be defined not *necessarily* these three, hough these are the ones that have most concerned us here: see reference 10.

are shown in Fig. 35 (a). The characteristics of such networks may be identical if the unbalanced one be measured by an unbalanced generator and receiving device (e.g. a valve voltmeter) and the balanced one by similar balanced means. However, these networks could not, in general, be interchanged as units in a chain of networks. But if the characteristics of an unbalanced network have been calculated, then these characteristics may also be applied to the equivalent balanced one.

A set of "universal characteristics," such as those in Fig. 33 (d) or (e), represents the behaviour of a whole family of networks, which have



*(a) Balanced and Unbalanced (Limited) Equivalent Networks*



*(b) The T and π Equivalent Networks for a Pair of Coils Coupled Mutually (Complete Equivalents)*

Fig. 35.—Illustrating Limited Equivalence and Complete Equivalence.

a fixed arrangement of elements but in which the values of the elements are scaled up or down. Such networks may be considered as potentially limited equivalents in the sense that they have similar forms of response which can be made identical by adjustment of the element values. Thus, taking the network in Fig. 31 (d) as an example, whose characteristics are those in Figs. 33 (d) and (e), we may increase $C$ to $nC$, reduce $L$ to $L/n$ and $R$ to $R/n$. This means that the horizontal relative frequency scale $\omega CR$ remains unchanged and so does the value of $L/CR^2(=k)$, so that the new network has identical shapes of steady-state response characteristics.

It was suggested in Sec. 5, Chapter 1, that the waveforms of the

normal modes of oscillation of simple meshes are unaltered if the values of $L$, $C$, and $R$ be changed in this way, but that the time scale only is affected. Thus the transient responses of simple meshes may be represented by "universal" transient characteristics, as exampled by those in Fig. 8. The same idea is applicable to more complicated networks if the scaling of the element values is the same for every mesh in the network.

A very useful example of complete equivalence is shown in Fig. 35 (b) which represents the $T$ and $\pi$ equivalent networks for a pair of coils coupled mutually. These circuits are equivalent at all frequencies and are interchangeable in an external circuit, since they have identical input, output and transfer impedances.

In brief, completely equivalent networks have identical characteristics in all respects, whereas networks possessing limited equivalence (according to our present definition) have certain similarities in their characteristics, so that measurements or calculations of the steady-state response of one network over a continuous range of frequencies may be applied to the others directly. In this way considerable labour may sometimes be saved, both as regards the duplication of response calculation and as regards the reduction of a given network to its simplest form, before attempting analysis. It is not proposed to deal here with all the forms of equivalent network structures, but only with certain basic equivalences which help in an understanding of network response forms. The general theory of network equivalence is dealt with by many standard books on communication networks.[10]

It may be shown, by comparison of the differential equations which represent the relations between currents and voltages, that certain networks have equations of the same mathematical form, term by term, though by appearance the networks may seem to be quite different. For example, it was pointed out in Sec. 2, Chapter 1, that the differential equations 3 and 4, representing the behaviour of the two networks in Fig. 1 (a) and (b) respectively, are identical in form, the constants $L$, $1/C$, $R$ in one equation being replaced by $C$, $1/L$, $1/R$ in the other. Also, the variables of current and voltage are interchanged. Since these equations are analogous in this respect it is to be expected that their solutions, representing the transient behaviour of the two networks, will be of the same form, but that one solution will give a current transient waveform and the other a voltage transient waveform, and that these waveforms will be identical.

This principle may be applied to any linear, *planar** network, and it may be shown that all such networks go in pairs [10] called *duals*, having identical forms of differential equations, but with current and voltage interchanged, as well as the constants $L$ and $C$, $R$ and $1/R$. The complete steady-state and transient characteristics of such networks are of the same form, with the provision that current in one network is replaced by voltage in the other and *vice-versa*. By a *planar* network we mean one which may have its circuit drawn on paper without any cross-over branches being required.†

There are simple rules [12] for forming the dual of a given network, originally due to Cauer.[(E)] Such rules may be arrived at by a consideration of how the various terms arise in the differential equations representing the behaviour of dual networks. For example, in these same equations, 3 and 4, it may be seen that the term giving the voltage drop across a series element in one network corresponds to the term giving the current in a shunt element in the other (dual) network; also the expression for the voltage across an inductance is similar to the expression for the current through a capacity, and likewise the expressions for current in an inductance and voltage across a capacity are similar. Again the expressions for voltage drop across a resistance corresponds to the current in a conductance, and *vice-versa*. The current round the loop, in equation 3 for circuit in Fig. 1 (a), corresponds to the voltage across the branches, in equation 4 for circuit in Fig. 1 (b).

Thus a series (shunt) inductance is the dual of a shunt (series) capacity, a series (shunt) resistance is the dual of a shunt (series) conductance, while a series (shunt) voltage generator is the dual of a shunt (series) current generator.

Extending the rules for these elementary networks to more complicated ones, possessing a number of meshes or branches, leads to a simple geometrical construction for the formation of a dual network, which we shall illustrate by the example of Fig. 36. Two sets of dual networks are given in this figure: (a) shows the same elementary tuned circuits that we have already examined in relation to their differential equations 3 and 4, while (b) shows a constant-$K$

* For application to non-planar networks see A. Block "On methods for the Construction of Networks Dual to Non-planar Networks," *Proc. Phy. Soc.*, Vol. LVIII, 1946, p. 677.

† For earliest reference, see A. Russell. Ref. 4, Chapter 2.

section with a resistance termination, and its dual. The dual networks are formed as follows:

(1) Inside every mesh mark a reference point ($A$, $B$, $C$ . . .) with one point outside the whole network ($D$). These form the branch points of the dual network.

(2) Join each point to its neighbour by a line passing through every element common to the two meshes surrounding the points. These lines represent the branches of the duals.

(3) Replace every element through which such a line passes by the dual of the element.

The network thus formed is the dual of the first network.

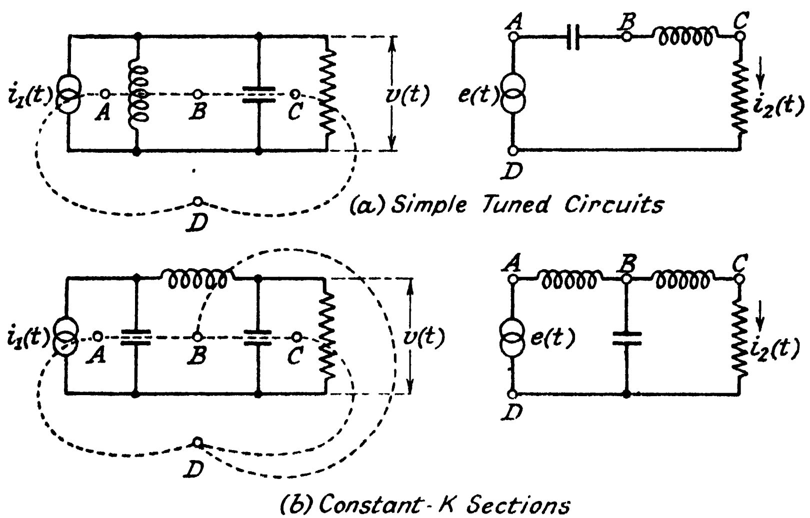



Fig. 36.—Rules for the Formation of Dual Networks.

This duality property may be applied in reverse, as the reader may test with these examples, by taking the right-hand networks in Fig. 36, marking suitable reference points as directed and forming the duals of these networks, which will be seen to be the same as the original networks. Thus if a network (I) is the dual of a network (II), then (II) is the dual of (I). The duals of certain networks are the same as the networks themselves (*self-duals*).

This dual property, as we have described, refers to a network *structure* only, and not to the values of the elements in the networks.

If the networks are to be *complete* duals, the elements must have their values adjusted so that corresponding reactances in the two dual structures have a constant product, as is discussed in the following section.* Then the characteristics of the networks will be identical, but with currents in one network proportional to voltages in the other. Thus if the waveforms $i_1(t)$ and $e(t)$ of the dual generators are identical, the waveforms of the network responses $v(t)$ and $i_2(t)$ are themselves identical; the impedance of a network equals the admittance of its complete dual between any pair of terminals.

We have so far not considered how to deal with mutual inductance, but this is readily fitted into the picture. When applying rule (2) above it must be remembered that every pair of coils coupled together really consists of three inductances—the self-inductance of each coil and the mutual inductance. Thus one line must be joined between the appropriate reference points for each of these inductances, and the dual of each will be a capacity. The same result would apply if the coupled coils be replaced by their equivalent circuit consisting of self inductances alone, as shown in Fig. 35 (b).

## 29. Foster's reactance theorem †—the characteristics of two-terminal reactance networks

It may at first appear that the characteristics of a network, containing a number of elements, can take on a large variety of forms, depending upon the arrangement of the elements. However, it may be shown that if we consider purely *reactive* networks, containing no resistances to dissipate energy, then considerable systematising of the characteristics is possible, and that if the number of elements in a 2-terminal network is known then there are only four possible forms which may be assumed by the driving-point impedance characteristics.

The behaviour of filters and other 4-terminal reactive networks may be understood more readily by a study of the individual characteristics of the 2-terminal arms of which such networks are composed. For example, the two basic arrangements of series and shunt arms in the formation of a filter section are the $T$ and $\pi$ structures, illustrated in Figs. 35 (a) and 31 (e) respectively, in which each arm, $Z_1$, $Z_2$, and $Z_3$, is a 2-terminal network.

In Chapter 1 we have seen that the presence of both an inductance $L$ and a capacity $C$ in the same mesh leads to "resonance,"

* See p. 101.

† Originally stated by G. A. Campbell, without proof. See reference (B).

and that the force-free response of such a mesh is a continuous oscillation, having a frequency $f_0$ where:

$$2\pi f_0 = \frac{1}{\sqrt{LC}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (139)$$

assuming that there is no resistance in the mesh. If a 2-terminal network contains a large number of separate inductances and capacities then combinations of these elements cause the network to have a number of separate resonant frequencies. Before attempting to find the number of such resonances or to determine the characteristics of such a reactive network in a general way, let us consider the simplest cases and build up from there.

The simplest form that a reactive arm may take consists of a single element, either an inductance $L$ or a capacity $C$, as in Fig. 37 (a). The reactances, which are plotted in this diagram on an $\omega$ scale, are of course:

$$\left.\begin{array}{l} X = j\omega L \\ X = 1/j\omega C \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (140)$$

These elements are duals, and their product is constant, independent of frequency. Denoting this constant by $R^2$,

$$(j\omega L)\left(\frac{1}{j\omega C}\right) = \frac{L}{C} = R^2 \quad . \quad . \quad . \quad . \quad (141)$$

The constant $R^2$ has the units of resistance (ohms) squared if $L$ and $C$ are given in henries and farads.

Fig. 37 (b) shows the only possible forms of two element reactive arms. These are series tuned (resonant) arm and shunt tuned (anti-resonant), both having zero resistance and hence infinite Q. The reactance diagram of the resonant arm crosses the zero reactance axis at a frequency $\omega_1$, and this point is referred to as *a zero*. The reactance diagram of the anti-resonant arm is discontinuous at the frequency $\omega_2$, going to infinite reactance at this point, which is referred to as *a pole.*

The reactances, as plotted, of these two arms are respectively:

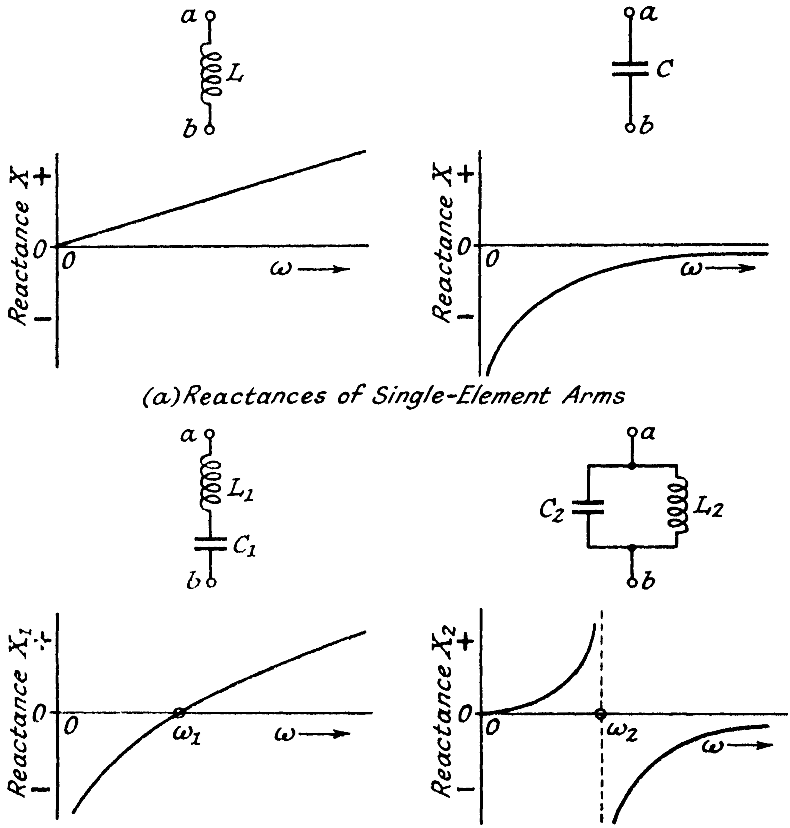
$$\left.\begin{array}{ll} X_1 = \dfrac{jL_1}{\omega}(\omega^2 - \omega_1^2) & \text{(for resonant arm)} \\[2ex] X_2 = \dfrac{\omega}{jC_2(\omega^2 - \omega_2^2)} & \text{(for anti-resonant arm)} \end{array}\right\} \quad . \quad . \quad (142)$$

where $\omega_1^2 = \dfrac{1}{L_1 C_1}$ and $\omega_2^2 = \dfrac{1}{L_2 C_2}$ . . . (143)

Again, these two arrangements of elements are duals. If corre-

sponding pairs of reactances in these duals are adjusted in value so that the product of each pair is constant, the arms become complete duals or *inverse* networks.* That is, if

$$\frac{L_1}{C_2}=\frac{L_2}{C_1}=R^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (144)$$




(b) Reactances of Two-Element Arms

Fig. 37.—Reactance Diagrams for Simple 2-Terminal Arms.

then the reactances in equation 142 become reciprocals at all frequencies, since $\omega_1=\omega_2$. Thus the product of these two reactances becomes a constant at all frequencies:

$$X_1X_2=\frac{L_1}{C_2} \text{ (from 142)}=R^2 \qquad (145)$$

* See p. 99.

and the zero of one arm $\omega_1$ corresponds in frequency to the pole of the other arm $\omega_2$.

If the case of three elements be considered, either two inductances and one capacity or two capacities and one inductance, it will be found that there are only four possible ways of connecting them in circuit between two terminals (excluding the cases in which elements of the same kind appear in series or parallel, which simplify to become 2-element arms). It will further be found that these four possible arrangements of elements, illustrated by Fig. 38, designated (*a*), (*b*), (*c*), and (*d*), fall into two groups (*a*), (*c*) and (*b*), (*d*), and that opposite members of each group (*a*), (*b*) and (*c*), (*d*) are dual networks. This becomes apparent by an inspection of these four possible arrangements, as the reader may test for himself, using the simple geometrical construction discussed in the last section. But it is not immediately obvious, though it may be proved,* that

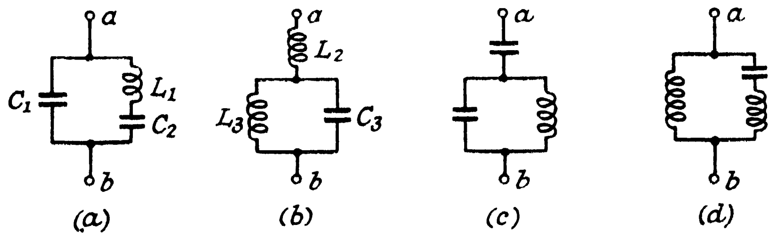



Fig. 38.—Possible Arrangements of 3-Element Arms.

members of the same group (*a*), (*c*) and (*b*), (*d*) are limited equivalents having the same form of impedance characteristic, and that these may be made complete equivalents by suitable adjustment of the element values, in which case their frequency characteristics become identical.

If the dual pairs (*a*), (*b*) and (*c*), (*d*) are to be made complete duals or inverse networks, then, as we have already seen for the 2-element arms, the reactance of corresponding elements in the duals must have a constant product. For example, in (*a*) and (*b*) of Fig. 38, if we make

$$\frac{L_1}{C_3}=\frac{L_2}{C_1}=\frac{L_3}{C_2}=R^2 \text{ (constant)} \qquad (146)$$

the arms will be truly inverse.

* See reference 10 and Chapter 5. A theorem known as the "L-type network equivalence theorem" makes it possible to determine all the equivalent forms of a two-terminal reactance arm containing a given number of non-redundant elements.

The reactance of the arm (*a*) is given by:

$$X_A=\frac{(1/j\omega C_1)(j\omega L_1+1/j\omega C_2)}{(1/j\omega C_1+j\omega L_1+1/j\omega C_2)}$$

which may be simplified to:

$$X_A=\frac{1}{j\omega}\cdot\frac{(1-\omega^2L_1C_2)}{(C_1+C_2-\omega^2L_1C_2C_1)}$$

This expression may then be written as:

$$X_A=\frac{-j}{\omega C_1}\cdot\frac{(\omega^2-\omega_1{}^2)}{(\omega^2-\omega_2{}^2)} \quad \text{. . . . . . (147)}$$

where $$\omega_1{}^2=\frac{1}{L_1C_2} \quad \text{. . . . . . . . . . (148)}$$

gives the frequency of resonance of the series elements $L_1$ and $C_2$ in this arm (*a*), and where:

$$\omega_2{}^2=\frac{C_1+C_2}{L_1\,.\,C_1C_2} \quad \text{. . . . . . (149)}$$

gives the anti-resonance of the element $L_1$ and the capacity elements $C_1$ and $C_2$, which are effectively in series.

In a similar manner it may be shown that the reactance of the arm (*b*) in Fig. 38, which is the dual of (*a*), is:

$$X_B=j\omega L_2\,.\,\frac{(\omega^2-\omega_4{}^2)}{(\omega^2-\omega_3{}^2)} \quad \text{. . . . . . (150)}$$

where $$\omega_3{}^2=\frac{1}{L_3C_3} \quad \text{. . . . . . . . . . (151)}$$

gives the anti-resonance of the elements $L_3$ and $C_3$, and where:

$$\omega_4{}^2=\frac{L_2+L_3}{C_3\,.\,L_2L_3} \quad \text{. . . . . . (152)}$$

gives the resonance of the element $C_3$ and the inductance elements $L_2$ and $L_3$, which are effectively in parallel (it must be remembered that the terminals *a* and *b* will be connected by some external path).

If the condition of equation 146 holds, the resonant frequency, or zero, of one arm will coincide with the anti-resonant frequency, or pole, of the other, and *vice versa*. That is:

$$\left.\begin{array}{l}\omega_1=\omega_3\\ \omega_2=\omega_4\end{array}\right\} \quad \text{. . . . . . . . . (153)}$$

in which case the product of the reactances of the arms (*a*) and (*b*) becomes constant. From 147 and 150:

$$X_A X_B = \frac{L_2}{C_1} \tag{154}$$

Fig. 39 shows these reactances $X_A$ and $X_B$ plotted against $\omega$; these diagrams also apply to the corresponding equivalent arms (*c*) and (*d*) in Fig. 38. It may be seen from an examination of the expressions for the reactance curves, in equations 142 and 143 for 2-element arms, and in 147 and 150 for 3-element arms, that the shapes of these reactance curves are completely determined by the values of $\omega_1$, $\omega_2$, etc. (the poles and zeros), together with one constant (an inductance or capacity value). Although in such a simple case the resonant and anti-resonant meshes may possibly be traced out and reactance expressions such as 147 and 150 derived

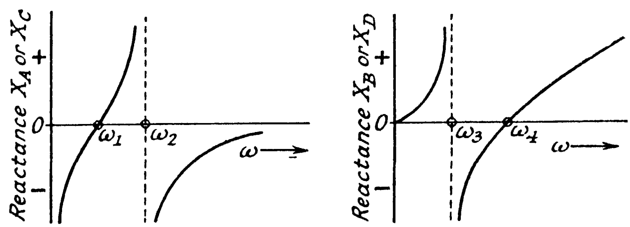



Fig. 39.—Reactances of the 3-Element Arms in Fig. 38.

either by inspection or by calculation, as we have done here, it is not readily possible to do this in the case of more complex arms. However, the forms of the reactance curves may be deduced, for an arm containing an irreducible number $N$ of elements (i.e. one containing no like elements in series or parallel) by a theorem usually attributed to R.M. Foster.[13] This theorem extends the results we have obtained for 2- and 3-element arms to arms containing any number of lumped reactance elements, and states that the driving-point reactance characteristic of an arm is determined completely by the location of its (finite)* poles and zeros on the frequency axis, together with the value of one constant.

The theorem further states that the reactance characteristic must assume one of four forms, as shown in Fig. 40, either of the forms (*a*) or (*b*) if the number $N$ of elements in the arm is even, and either

* i.e. not including the pole or zero at $\omega=0$ or $\infty$.

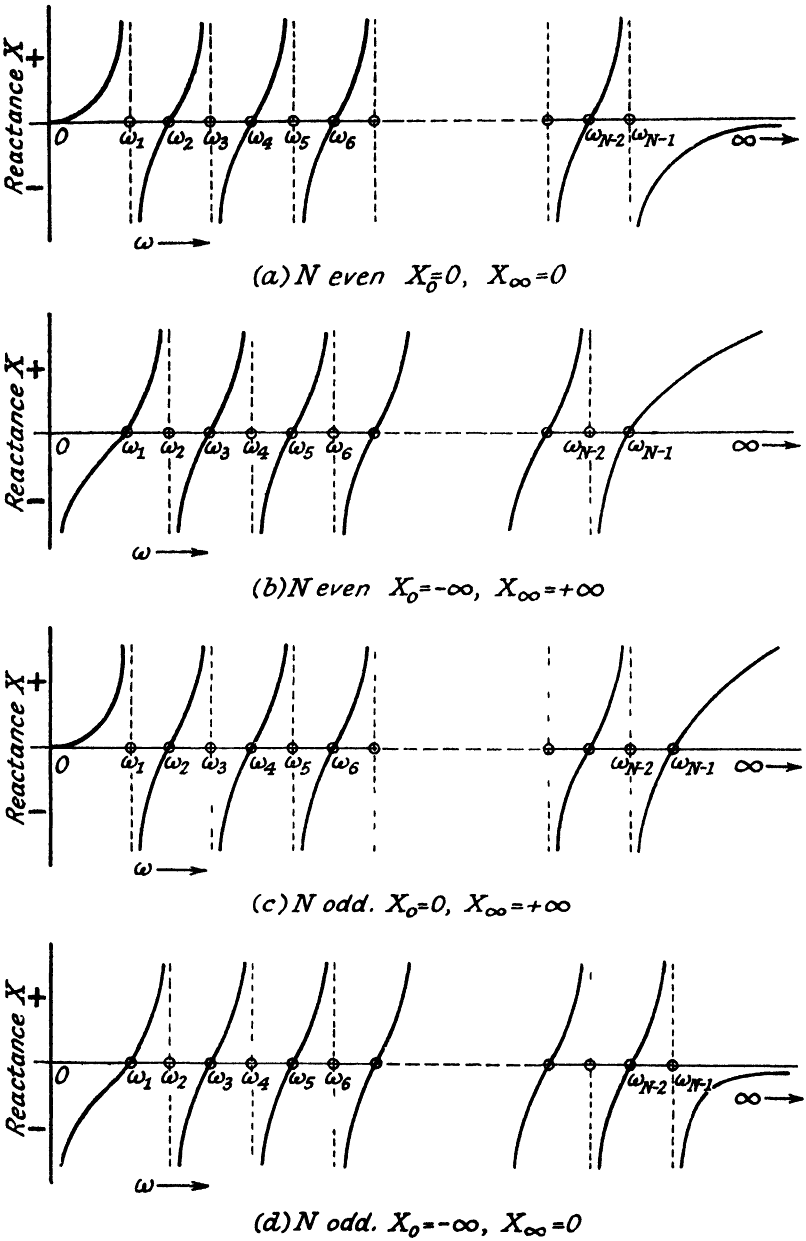



Fig. 40.—The Four Possible Forms of Reactance Characteristic of a 2-Terminal Arm Composed Entirely of Reactance Elements.

of the forms (*c*) or (*d*) if $N$ is odd. As the number of elements $N$ is increased, the number of possible arrangements of these elements between the two terminals increases very rapidly, but however many arrangements there may be, for *a given number* $N$ the reactance characteristics must fall into one of two groups. All the members of one group must be limited equivalents and must be duals to members of the other group. The characteristics of one group will be of the form (*a*) and of the other group of the form (*b*) if $N$ is even. These characteristics are reciprocal. Similarly if $N$ is odd, the characteristics of the two groups will be of the form (*c*) and (*d*) respectively, which are also reciprocals.

The total number of poles and zeros in every case will be one less than the number of elements $N$, even or odd, *not including* poles or zeros at $\omega=0$ or $\infty$. The slope of these reactance/frequency characteristics, $dX/d\omega$, is everywhere positive, and the poles and zeros must therefore alternate.

It has been shown [14] that the analytical expressions for these reactance characteristics are as follows.

For the cases in which $X=0$ at $\omega=0$ (Fig. 40 (a) and (c)):

$$X=j\omega H\,.\,\frac{(\omega^2-\omega_2{}^2)(\omega^2-\omega_4{}^2)(\omega^2-\omega_6{}^2)\;.\;.\;.}{(\omega^2-\omega_1{}^2)(\omega^2-\omega_3{}^2)(\omega^2-\omega_5{}^2)\;.\;.\;.}\quad .\;.\quad (155)$$

while for the cases in which $X=-\infty$ at $\omega=0$ (Fig. 40 (b) and (d)):

$$X=-j\frac{H}{\omega}\,.\,\frac{(\omega^2-\omega_1{}^2)(\omega^2-\omega_3{}^2)(\omega^2-\omega_5{}^2)\;.\;.\;.}{(\omega^2-\omega_2{}^2)(\omega^2-\omega_4{}^2)(\omega^2-\omega_6{}^2)\;.\;.\;.}\quad .\;.\quad (156)$$

where $H$ is a constant which may have the units of $L$ or $1/C$. The factors in the numerators give the frequencies of the zeros and those in the denominators the frequencies of the poles. The total number of factors is one less than $N$, the number of reactive elements, in both 155 and 156.

The driving-point reactances given by these two equations will be exact reciprocals if the poles of one correspond exactly in their frequencies to the zeros of the other (of course, if $N$ be the same for both reactances), and the arms will be inverse networks. In Fig. 40 the spacing between adjacent poles and zeros is shown uniform; this is merely for convenience, by way of illustration.

It will be observed that the reactance diagrams on Figs. 37 and 39, which were obtained for the simple cases of 2- and 3-element arms, are special cases of the general reactance diagrams in Fig. 40. Similarly, the reactance expressions 142, 147, and 150 are special cases of the general reactance expressions 155 and 156.

If a 2-terminal arm contains an irreducible even number, $N$, of elements, there will be an equal number of $L$ and $C$ elements; if the number $N$ is odd, there will be one extra $L$ or $C$ element. It is always possible to derive an equivalent circuit for an arm, containing $N$ reactive elements, in two of the four simple circuit configurations shown in Fig. 41, either as in (a) and (b) if $N$ be even or as in (c) and (d) if $N$ be odd.* These simple forms of circuit, consisting of a chain of resonant or anti-resonant arms, are known as the *canonic*

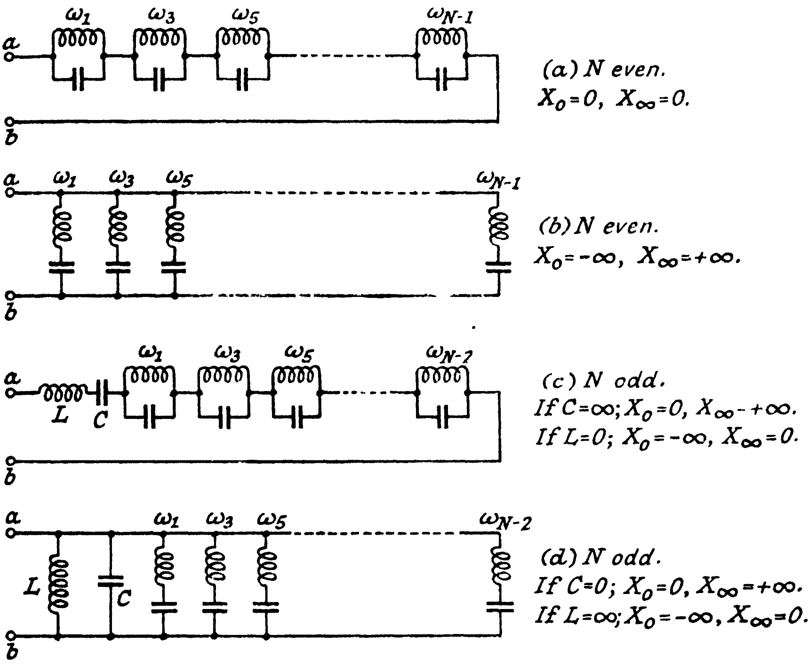



Fig. 41.—The Four Canonic Forms of 2-Terminal Reactance Arms, having the Characteristics of Fig. 40.
[If $N$ is odd, as for (c) and (d), either $C$ or $L$ must be absent.]

*forms*, and provide a simple interpretation of the four basic reactance diagrams in Fig. 40, since the resonant arms and anti-resonant meshes are more obvious than in any other element arrangements. The resonant and anti-resonant frequencies of each tuned circuit in these networks, corresponding to the zeros and poles in the reactance diagrams, are indicated in Fig. 41 as $\omega_1$, $\omega_3$, $\omega_5$, etc.

The arms in Fig. 41 (a) and (b) having an even (irreducible) number

* See footnote, p. 102.

of elements are, of course, dual networks, as the reader may test for himself by applying the graphical rule explained in Sec. 28, and become exact duals (or inverse networks) if the poles of one, $\omega_1$, $\omega_3$, $\omega_5$ . . ., etc., correspond to the zeros of the other. Similarly the arms (c) and (d) are inverse for the case of $N$ being odd. It must be remembered that the terminals *ab* will be closed by some external path if the reactance of the network is to have any meaning.

It is occasionally very useful to be able to find the number of modes of oscillation of a reactive arm, and this is decided by the number of poles and zeros in the "Foster reactance diagram" of the arm. If the arm be reduced to its canonic form by successive applications of the L-transformation theorem,* the frequencies of the oscillations may be determined by inspection. It is important that the arm be reduced to the minimum number of elements, and if mutual inductance is present this may be included by replacing the coupled inductances by their $T$ or $\pi$ equivalents (Fig. 35 (b)). The basic form of the reactance diagram may always be determined once the number of elements has been made irreducible. Thus the values of $X$ at $\omega=0$ and $\omega=\infty$ may be found by inspection and the number of elements $N$ counted, which gives the number $(N-1)$ of poles and zeros.

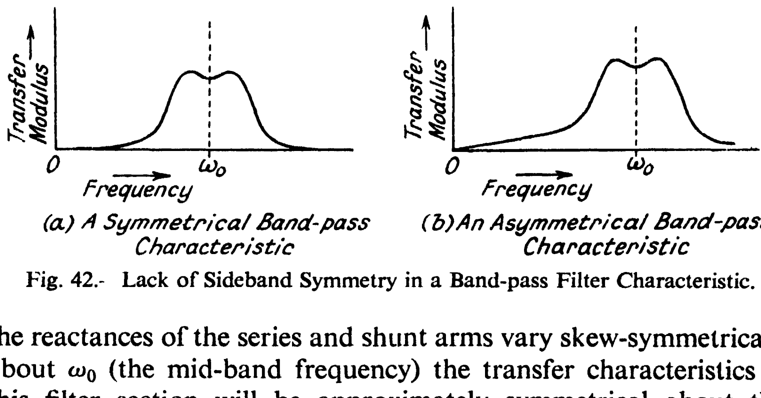
## 30. Symmetric and asymmetric characteristics

There is an important classification of network characteristics that should be studied, which refers to the degree of symmetry possessed by the characteristics of band-pass filters, and by other similar structures normally designed to transmit amplitude-modulated waves. For example, Fig. 42 illustrates the transfer characteristics of two common types of band-pass filter (only the moduli are shown). The characteristic of the structure (a) is seen to be reasonably symmetrical about its mid-band frequency $\omega_0$, while that of the structure (b) is markedly asymmetrical, and it will be shown later that, in the case (a), the characteristic becomes more and more exactly symmetrical about $\omega_0$ as the ratio of bandwidth: mid-band-frequency becomes smaller.

The series or shunt arms (of reactive elements) of which the $T$ or $\pi$ sections of any filter are composed must have characteristics of the general forms shown in Fig. 40. If the arms have an even number of elements the reactance characteristics, (a) and (b), are,

* See footnote, p. 102.

broadly speaking, of a symmetrical type, in that they have either a pole or a zero at both zero and infinite frequency. If the number of elements be odd, asymmetry is introduced because there will be a zero at $\omega=0$ and a pole at $\omega=\infty$ or vice versa. It is perhaps necessary to qualify the use of the term "symmetry" here, since that implies some central axis. The symmetry with which we are dealing is more obvious when we consider simple reactive arms of a few elements, as are most commonly used in practical filters. For example, the series and shunt arms of the $\pi$ section of Fig. 43 (c) have reactance characteristics of the forms shown in Fig. 37 (b) for 2-element arms. These characteristics are *approximately* (skew) symmetrical about $\omega_1$ and $\omega_2$ respectively, and if these frequencies be made identical, at $\omega_0$, the arms become true inverse networks, in which case the filter section becomes a constant-$K$ section. If




(a) A Symmetrical Band-pass Characteristic (b) An Asymmetrical Band-pass Characteristic

Fig. 42.- Lack of Sideband Symmetry in a Band-pass Filter Characteristic.

the reactances of the series and shunt arms vary skew-symmetrically about $\omega_0$ (the mid-band frequency) the transfer characteristics of this filter section will be approximately symmetrical about this frequency. (As in Fig. 42 (a)).

This is proved by writing the transfer impedance characteristic in terms of the series and shunt arm reactances $X_1$, $X_2$, and $X_3$. In the general cases, as illustrated by the block diagrams of Fig. 43 (a) and (b), we have:

$$\left.\begin{aligned} v&=i_1X_3 \\ i&=i_1+i_2 \\ \frac{i_1}{i_2}&=\frac{X_2}{(X_1+X_3)} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \qquad (157)$$

Eliminating $i_1$ and $i_2$ gives for the transfer impedance:

$$\frac{v}{i}=\frac{X_2X_3}{(X_1+X_2+X_3)} \quad . \quad . \quad . \quad . \qquad (158)$$

If the arms all have an even number of elements, the values of $X_1$, $X_2$, and $X_3$ at any frequency above $\omega_0$ (say $\omega_0+\omega$) are positive, and these reactances will have *nearly* equal negative values $-X_1$, $-X_2$, and $-X_3$ at the frequency $(\omega_0-\omega)$, below $\omega_0$. But the value of $v/i$ in 158 is unaltered, though reversed in sign, when $X_1$, $X_2$, and $X_3$ become negative. Thus this transfer impedance has a symmetrical modulus and a skew-symmetrical phase characteristic about the mid-band frequency $\omega_0$. This symmetry is not exact, but

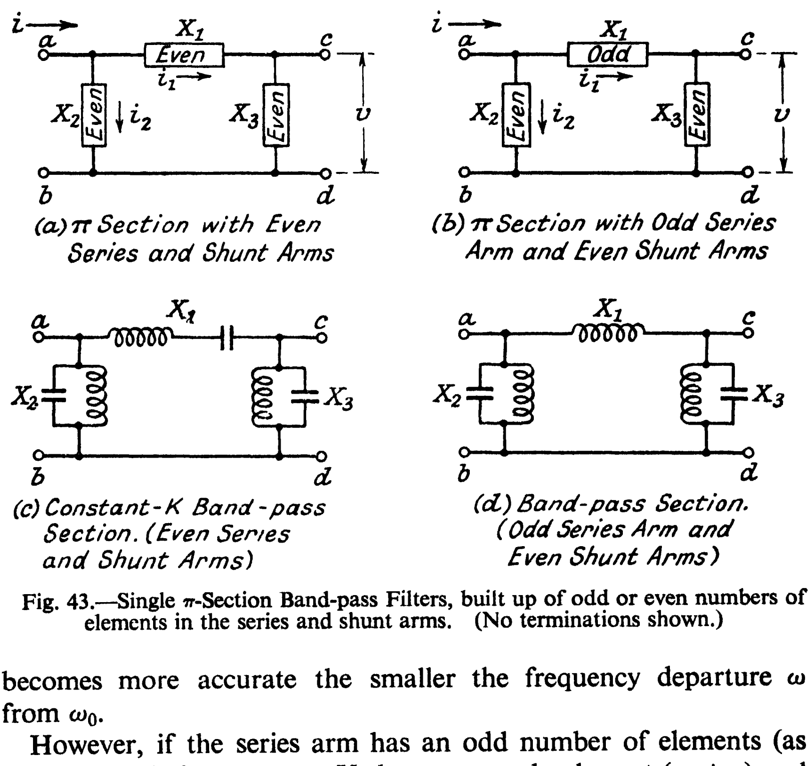



Fig. 43.—Single $\pi$-Section Band-pass Filters, built up of odd or even numbers of elements in the series and shunt arms. (No terminations shown.)

becomes more accurate the smaller the frequency departure $\omega$ from $\omega_0$.

However, if the series arm has an odd number of elements (as in Fig. 43 (b)) its reactance $X_1$ has not equal values at $(\omega_0+\omega)$ and $(\omega_0-\omega)$, so that the expression for $v/i$ is not symmetrical about $\omega_0$.

For example, the transfer impedance of the $\pi$ section shown in Fig. 43 (d), in which the series arm has the reactance characteristic shown in Fig. 37 (a), does not vary symmetrically about any frequency. Thus, although the shunt arms may vary symmetrically about $\omega_0$, the series arm does not, and the resulting transfer characteristic of the section cannot be symmetrical about the mid-band frequency $\omega_0$.

The near symmetry of the first characteristic, Fig. 42 (a), is typical of constant-$K$ filter structures in which both the series and shunt arms have an even number of elements, while the essential asymmetry of the second characteristic, Fig. 42 (b), indicates a filter structure in which the series and shunt arms contain an even and odd number of elements respectively, or vice versa, as illustrated in Fig. 43 (b) Also, if both the arms contain an even number of elements but are *not* true inverse networks, the structure is not constant-$K$, and will not have a near-symmetrical characteristic.

We have used the expression "near-symmetrical" here with a good reason. The reactance diagrams of Fig. 37 (b) (and others) are not exactly symmetrical about a centre frequency, since the frequency scale extends from zero to infinity, but they may be truly symmetrical, as geometric curves, if drawn on a logarithmic scale of frequency. The slight asymmetry of the frequency characteristic of a filter, which appears owing to this factor, may be called *inherent asymmetry*. It should be appreciated that all low-pass filters may be regarded as band-pass filters with zero frequency mid-band ($\omega_0=0$) and their characteristics, drawn in conjugate form as in Fig. 33 (d), are exactly symmetrical about zero frequency.*

## 31. Symmetry of filter characteristics on a logarithmic frequency scale

The elimination of this slight *inherent* asymmetry of the characteristics of a band-pass constant-$K$ filter, by drawing on a logarithmic scale of frequency, is best shown by the example of a single tuned circuit. This is the simplest form that the arms of such a filter may take (Fig. 43 (c)).

The characteristics of resistanceless resonant and anti-resonant arms on a linear frequency scale are illustrated by Fig. 37 (b). These are given by equation 142, and if we write $\omega_0$ for both the resonant frequency $\omega_1$ and anti-resonant frequency $\omega_2$, these become:

$$\left.\begin{aligned} X_1&=\frac{jL_1}{\omega}\,.\,(\omega^2-\omega_0^2) \quad \text{(for resonant arm)} \\ X_2&=\frac{\omega}{jC_2(\omega^2-\omega_0^2)} \quad \text{(for anti-resonant arm)} \end{aligned}\right\} \quad . \ . \quad (159)$$

* See Chapter 3, Sec. 25.

These characteristics are drawn on a logarithmic scale of frequency, in terms of $\omega/\omega_0$, in Fig. 44 (a) and (b). On such a scale, any linear distance $x$ from a chosen origin (where $\omega=\omega_0$) is given by:

$$x = \log_\epsilon \omega/\omega_0 \qquad \ldots\ldots \quad (160)$$

so that we may write:

$$\omega = \omega_0 \epsilon^x \qquad \ldots\ldots\ldots \quad (161)$$

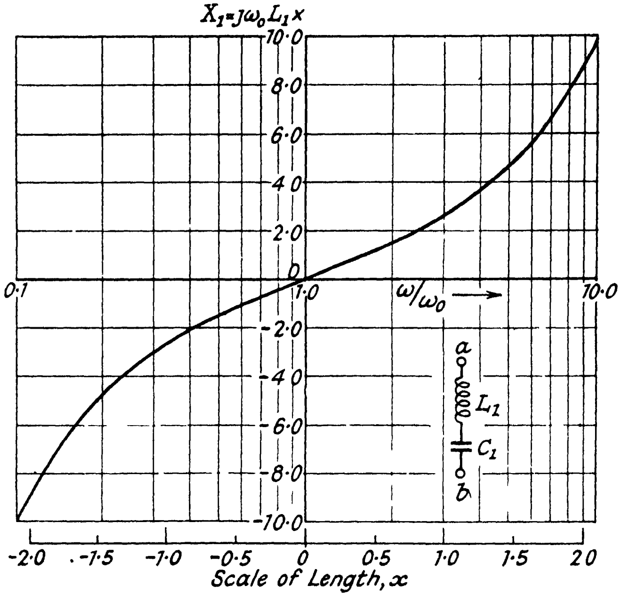



Fig. 44 (a).—Reactance Characteristic of Resonant Arm, on a Logarithmic Scale of Frequency.

for frequency, in equation 159, giving:

$$\left.\begin{aligned} X_1 &= j\omega_0 L_1(\epsilon^x - \epsilon^{-x}) \quad \text{(for resonant arm)} \\ X_2 &= \frac{1}{j\omega_0 C_2(\epsilon^x - \epsilon^{-x})} \quad \text{(for anti-resonant arm)} \end{aligned}\right\} \quad . \quad (162)$$

Both these expressions are quite (skew) symmetrical with respect to $x$, and the reactance characteristic curves in Fig. 44 are seen to be geometrically (skew) symmetrical.

Lack of perfect symmetry in filter characteristics may not be important in practice if it is small, but when it becomes large, as may happen with relatively wide-band filters, it means that the side-bands of a modulated carrier wave, tuned to the mid-band

frequency $\omega_0$, will be distorted asymmetrically. The effects of such distortion can be of practical importance, as will be discussed more fully in Chapter 7.

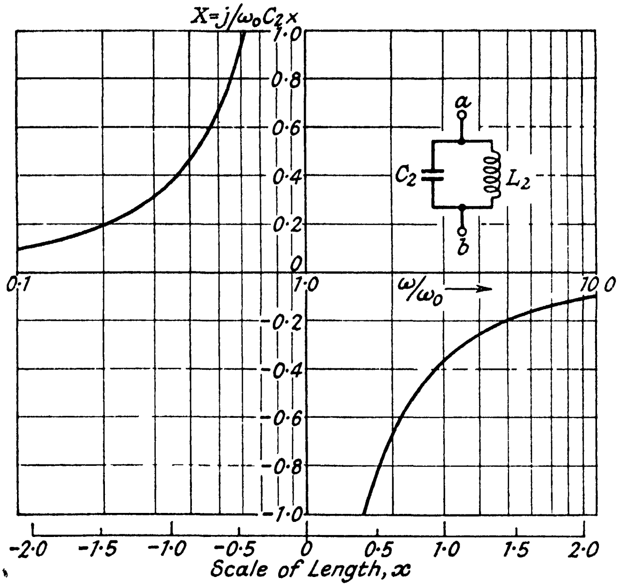



Fig. 44 (b).—Reactance Characteristic of Anti-resonant Arm, on a Logarithmic Scale of Frequency.

## 32. Some effects of dissipation on steady-state characteristics

The design of filter structures is based on completely loss-free elements, and their frequency characteristics are idealised in that the effects of dissipation are neglected. The forms of the frequency (transfer) characteristics may be derived from a knowledge of the characteristics of the individual series and shunt arms, in which respect Foster's reactance theorem is of great value. This theorem, as has been given here, is strictly true only for loss-free reactive arms.

If, in practice, the dissipation in the various elements has reasonable values, the theorem will be true qualitatively, in that the general form of the *impedance* characteristic of an arm containing a number of inductances and condensers will show a corresponding number of resonances and anti-resonances alternating with one another along a scale of frequency. There will be no *infinite* peaks of reactance, as in the cases of zero dissipation we have considered

(Figs. 37, 39, 40) and the frequencies of the resonances and anti-resonances may not correspond exactly with the zeros and poles of the dissipationless case.

The degree to which a characteristic departs from the ideal, dependent on the relative loss in the elements, is normally estimated by the values of Q for the various elements [15] in the circuit. We have already discussed, in Chapter 1, Sec. 5, the use of the factor Q as relating the energy-storage:energy-dissipation qualities of a single element or a tuned circuit, and have shown the effect, on the force-free oscillation of a tuned circuit, of varying its Q value (see

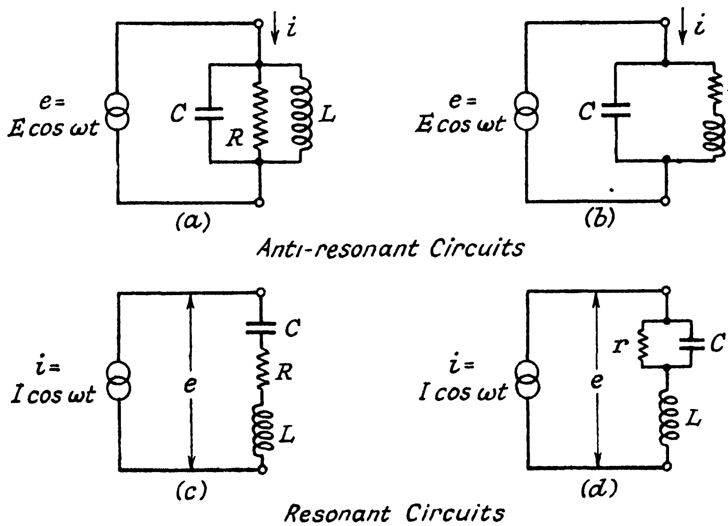



Fig. 45.—Circuits with Dissipation, (a) and (b) Anti-resonant Circuits and (c) and (d) their Resonant Duals.

Fig. 8). The effect on the steady-state frequency characteristics of a circuit is equally important, and a set of universal resonance curves for a single-tuned circuit is useful in this connection.

Fig. 45 shows four common types of lossy tuned circuit, the circuits (*c*) and (*d*) being the duals of (*a*) and (*b*) (as regards *structure*) respectively.

*Circuit* (*a*)

The admittance of this anti-resonant circuit $1/Z_a$ is given by:

$$\frac{1}{Z_a}=\frac{1}{R}+j\omega C+\frac{1}{j\omega L} \tag{163}$$

or $$Z_a = \frac{1/R - j(\omega C - 1/\omega L)}{(1/R)^2 + (\omega C - 1/\omega L)^2} = [C + jS] \quad . \quad . \quad (164)$$

giving the real and imaginary parts of the impedance $Z_a$. The amplitude and phase-shift characteristics are then:

Amplitude $$|Z_a| = \frac{R}{\sqrt{[1 + Q^2(\omega/\omega_0 - \omega_0/\omega)^2]}} \quad . \quad . \quad (165)$$

Phase-shift $$\tan \phi_a = -Q\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right) \quad . \quad . \quad . \quad . \quad . \quad (166)$$

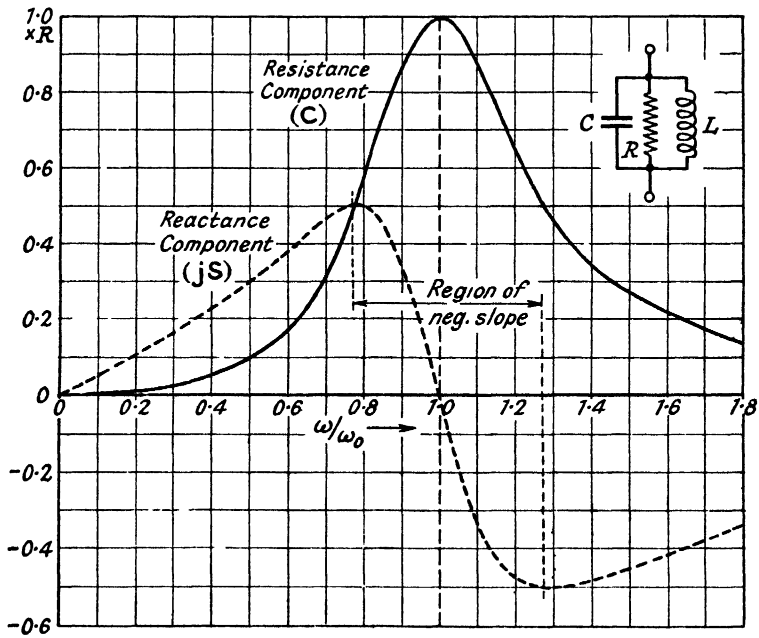



Fig. 46.—Resistance and Reactance Components of $Z_a$, the Impedance of the Anti-resonant Circuit of Fig. 45 (a), for Q=2·0.

Where: $$\omega_0{}^2 = \frac{1}{LC} \quad \text{and} \quad Q = \omega_0 CR \quad . \quad . \quad . \quad (167)$$

It may be seen that $\omega_0 CR$ is the Q value of this tuned circuit, from the definition of Q given in Chapter 1 (see p. 25), since:

$$(\omega_0 CR)^2 = \frac{CR^2}{L} = Q^2 \quad . \quad . \quad . \quad . \quad . \quad (168)$$

These equations, 165 and 166, represent also $1/Z_c$, the admittance (amplitude and phase) of the dual network (c).

Fig. 46 shows a plot of $Z_a(=1/Z_c)$ in terms of its real and imaginary parts, $C+jS$, for the particular value Q=2·0 at the resonant frequency, as an illustration of the form of these resistance and reactance component characteristics. The reactive component is seen to be of the same general shape as the reactance of the completely dissipationless circuit, sketched in Fig. 37 (b), even with this low chosen value of Q,* except that it does not go to infinity at the tuning point, but remains finite and has a negative slope, $dX/d\omega$, in this region. Dissipation, however small in practice, always reduces these infinite

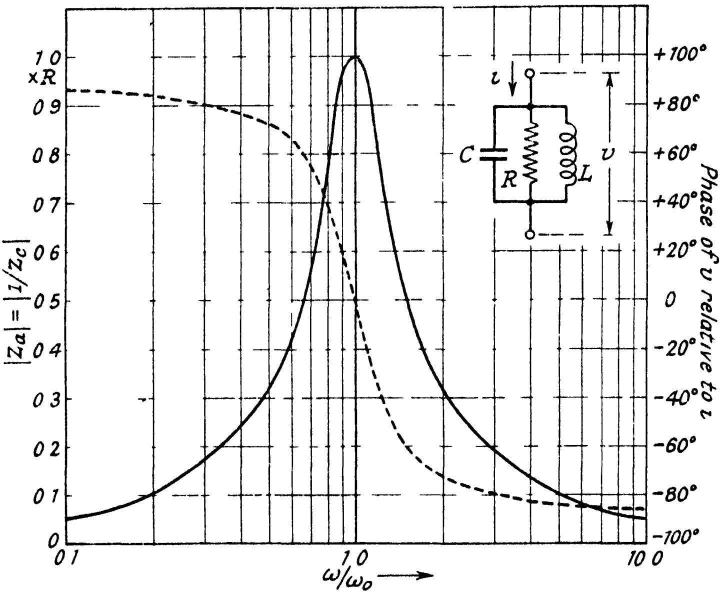



Fig 47 —Amplitude and Phase Characteristics of Damped Shunt Tuned Circuit, showing Symmetry on Logarithmic Scale.

reactances to finite values and necessitates a negative $dX/d\omega$ in the regions of anti-resonance.

If the amplitude and phase-shift characteristics (165) and (166) be plotted on a logarithmic scale of frequency, they will appear as symmetric curves, as may be seen by substituting $\omega=\omega_0\epsilon^x$ in equations 165 and 166, as explained in the preceding Section, 31.

These characteristics are shown in Fig. 47 for the value of Q=2·0, and illustrate the symmetry about $\omega_0$. The higher the circuit Q

* Such very low Q values occur commonly in television circuits, and in others necessitating relatively wide frequency bands.

value, the "sharper" the characteristic amplitude curve; there is a simple relation between the band-width, $\Delta\omega$, of this amplitude characteristic measured at $1/\sqrt{2}$ of the maximum amplitude at $\omega=\omega_0$ and the Q value. Putting $|Z_a|=R/\sqrt{2}$ in equation 165, gives:

$$1+Q^2\left(\frac{\omega}{\omega_0}-\frac{\omega_0}{\omega}\right)^2=2$$

or

$$\omega^2-\frac{\omega_0}{Q}.\omega-\omega_0{}^2=0$$

The difference of the roots gives the "band-width":

$$(\omega_1-\omega_2)=\Delta\omega=\omega_0/Q \quad . \quad . \quad . \quad . \quad . \quad (169)$$

This is a useful rule for estimating the Q value from the steady-state characteristics which is correct for all values of Q, however low, with this particular form of circuit.

*Circuit (b)*

The impedance of the damped tuned circuit of Fig. 45 (b), having resistance inserted in a non-symmetrical manner as shown, is given by:

$$Z_b=\frac{(r+j\omega L).1/j\omega C}{r+j\omega L+1/j\omega C}$$

which may be written as amplitude and phase characteristics:

$$\left.\begin{aligned} |Z_b| &= r\sqrt{\left[\frac{1+Q^2.\omega^2/\omega_0{}^2}{(1-\omega^2/\omega_0{}^2)^2+(1/Q^2)\omega^2/\omega_0{}^2}\right]} \\ \tan\phi_b &= \frac{1}{Q}\cdot\frac{\omega}{\omega_0}\cdot\left[Q^2\left(1-\frac{\omega^2}{\omega_0{}^2}\right)-1\right] \end{aligned}\right\} . \quad (170)$$

where:

$$Q=\frac{1}{\omega_0 Cr} \quad \text{and} \quad \omega_0{}^2=\frac{1}{LC} \quad . \quad . \quad . \quad . \quad . \quad (171)$$

It will be observed that the Q of this circuit (b) is the inverse of the Q of circuit (a), equation 167. This is so because the *L*, *C*, *R* elements in (a) are in parallel, whereas in (b) the *L*, *C*, *r* elements are in series with one another.*

The equations 170 represent also the admittance of the circuit (d), which is the dual of (b). It may be seen that neither the amplitude

* A quick test of whether the expression for Q is the right way up in any case of a damped tuned circuit is to put the resistance equal to zero and see whether Q→∞ or zero.

nor the phase-shift characteristics, given by equation 170, are symmetrical about $\omega/\omega_0=1$, in this instance, in contrast to the circuits (a) and (c). Only if the resistance be introduced into the tuned circuit in a symmetrical manner, as in (a) and (c), will the "near-symmetry" of the characteristics (on a normal linear frequency scale) be preserved. The curves in Fig. 48 give the amplitude and phase-shift characteristics for the special case of Q=2·0, and illustrate this lack of symmetry. The higher the Q value the better will the symmetry be, on a logarithmic scale of frequency, until in

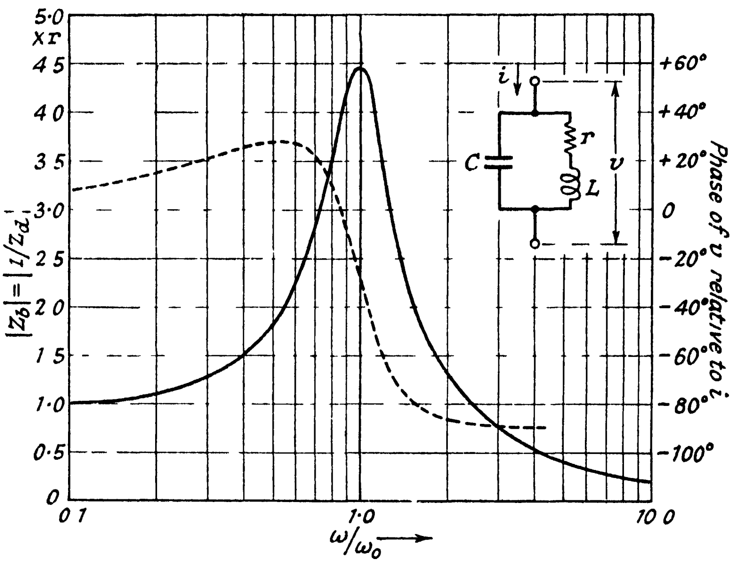



Fig. 48 —Amplitude and Phase Characteristics of Non-symmetrical Damped Tuned Circuit Q=2 0.

the limit, as $Q \to \infty$ the characteristics of both types of circuit (Figs. 47, 48) will become those of an undamped tuned circuit, Fig. 44.

## 33. Conversion of low-pass filters to equivalent band-pass filters

In Sec. 16 (Chapter 2) the equivalence was shown between an envelope wave and a carrier wave modulated in amplitude by this envelope, in that the spectrum of the envelope wave centred around zero frequency, while the spectrum of the modulated wave was identical but centred around the carrier frequency (see Figs. 18 and 19).

The frequency characteristics of networks are similar to wave spectra in that they are plots of amplitude and phase angle against frequency—they are essential continuous curves and therefore similar only to continuous spectra, such as the spectra of transient waves. Furthermore, the characteristics of low-pass networks centre around zero frequency (see Fig. 33) when plotted as conjugate components, and in the last two Sections it has been shown that the characteristics of certain band-pass networks are symmetrical about a mid-band frequency, apart from inherent asymmetry due to the use of a linear frequency scale.

These points suggest that there may be some connection between low-pass and band-pass (or high-pass and band-elimination)

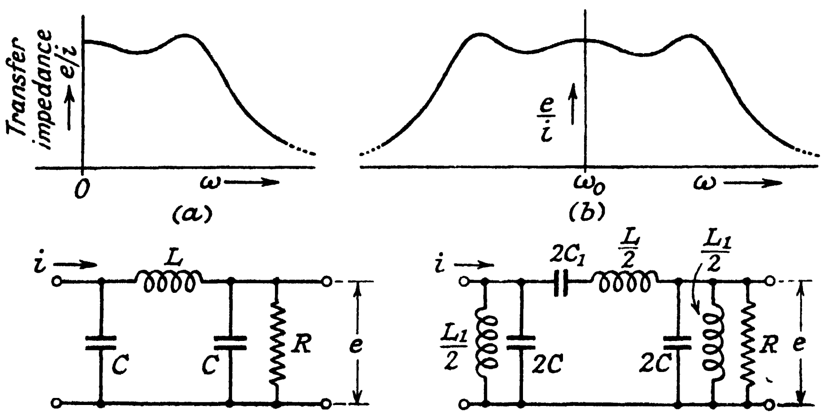



Fig. 49.—The Low-pass, Band-pass Analogy.

filters which have identical characteristics, centred around zero frequency and around some carrier frequency respectively. This relation is known, sometimes, as the "low-pass: band-pass analogy." [17]

Fig. 49 illustrates the theorem: (a) shows a low-pass filter section and its characteristic (amplitude only shown) and (b) the analogous band-pass filter section and its characteristic, symmetrical about a carrier frequency $\omega_0$. This analogous band-pass circuit is derived by adding a capacity in series with every inductance and adding an inductance in parallel with every capacity in the original low-pass circuit. Resistance elements have nothing added. The values of the added elements are such that every resonant or anti-resonant

arm so formed tunes to the same frequency, $\omega_0$. Thus in the example, Fig. 49:

$$\frac{1}{LC_1}=\frac{1}{L_1C}=\omega_0^2 \qquad . \quad . \quad . \quad . \quad . \quad (172)$$

The band-pass "analogue" thus formed will have a total band-width (i.e. including *both* side-bands) equal to the width of the original low-pass circuit, unless the inductance values be halved and the capacities doubled, still preserving the resonant frequencies of all the arms at $\omega_0$. Resistance values should be unaltered. The reader is reminded of the forms of the natural, force-free responses of such analogous circuits, which were examined in Sec. 6, Chapter 1.

For proof, it is sufficient to consider the relation between a single $L$ or $C$ element and its corresponding resonant or anti-resonant analogue. We have shown in the preceding Section that damped tuned circuits of the symmetrical form shown in Fig. 45 (a) and (c) have characteristics symmetrical about the tuning frequency, $\omega_0$, and only such circuits can possibly have their characteristics "transferred to zero frequency." It is also clear that circuits of the form in Fig. 45 (b) and (d) (which have been shown to have completely non-symmetrical characteristics) cannot, from their structure, be derived from the addition of an inductance or a capacity to a single existing element, according to the rules given above.

Using these rules enables us to form a band-pass "analogue" to any linear network whatever, but the converse is not necessarily true. Only those networks consisting entirely of tuned circuit branches of the form of Fig. 45 (a) or (c) (though their Q values may be infinite) and tuned to the same frequency $\omega_0$, can have low-pass "analogues."

Fig. 50 shows these resonant and anti-resonant arms together with their low-pass "analogues." Consider the series resonant arm.

Impedance of $r$, $L/2$, and $2C$ in series

$$=r+j\left(\frac{\omega L}{2}-\frac{1}{2\omega C}\right) \qquad . \quad . \quad . \quad . \quad . \quad (173)$$

Impedance of the low-pass analogue, $r$ and $L$ in series

$$=r+j\omega L \qquad . \quad . \quad . \quad . \quad . \quad . \quad (174)$$

These analogous arms have identical real components, $r$, independent of frequency. Let us see whether the reactive components can be made similar, but symmetrical about zero frequency and $\omega_0$ respectively. These reactances are also plotted on Fig. 50.

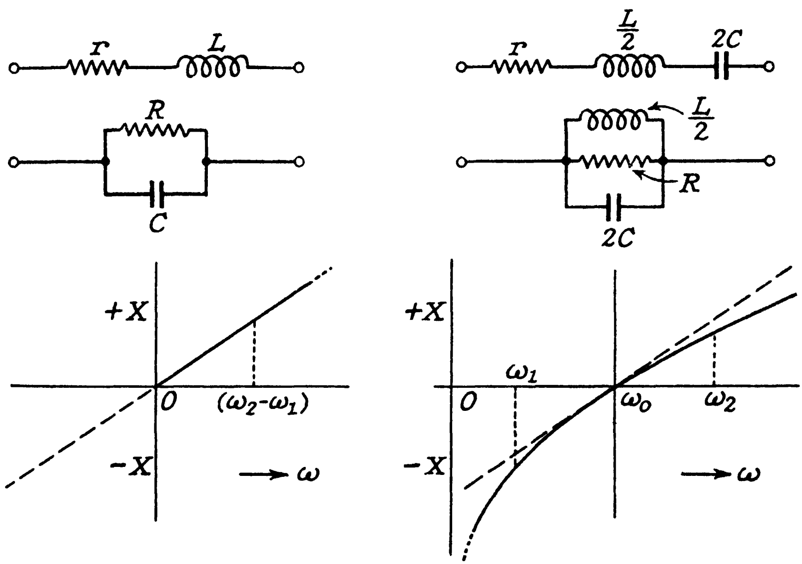
Let $\omega_1$ and $\omega_2$ be two frequencies at which the reactances $X$ of the resonant circuit are equal and of opposite sign:

$$X = -\frac{j}{2}\left(\omega_1 L - \frac{1}{\omega_1 C}\right) = +\frac{j}{2}\left(\omega_2 L - \frac{1}{\omega_2 C}\right) \quad . \quad (175)$$

Putting in $\omega_0^2 = 1/LC$ gives:

$$\left.\begin{aligned} 2X &= -j\omega_0 L\left(\frac{\omega_1}{\omega_0} - \frac{\omega_0}{\omega_1}\right) = j\omega_0 L\left(\frac{\omega_2}{\omega_0} - \frac{\omega_0}{\omega_2}\right) \\ \text{which simplifies to} \quad & \omega_0^2 = \omega_1\omega_2 \end{aligned}\right\} \quad . \quad (176)$$

Thus $\omega_0$ is the geometric mean of $\omega_1$ and $\omega_2$.




(a) Single Reactance Arms (b) "Band-pass Analogues"

Fig. 50.—The Low-pass: Band-pass Analogy, Applied to Simple Arms.

Substituting $\omega_1 = \omega_0^2/\omega_2$ in 176 gives:

$$2X = -j\omega_0 L\left(\frac{\omega_1}{\omega_0} - \frac{\omega_0}{\omega_0^2/\omega_2}\right) \quad \text{or} \quad X = j\frac{(\omega_2 - \omega_1)}{2}L \quad . \quad (177)$$

But this is simply the reactance of an inductance $L/2$ at the frequency $(\omega_2 - \omega_1)$, that is, the reactance shown in Fig. 50 (a), which is the result of "transferring $\omega_0$ to zero frequency." The whole characteristic may, strictly speaking, be transferred to zero frequency

provided that $\omega_0$ is the arithmetic mean of $\omega_1$ and $\omega_2$ instead of the geometric mean as given by 176. The result of making such an assumption would be a linear reactance characteristic for the tuned circuit, as shown by the dotted line, Fig. 50 (b). This assumption is justified in most practical cases when we are concerned with small values of $(\omega_2 - \omega_1)$ compared with $\omega_0$, that is to say in cases of relatively narrow band-width.

This same proof applies also to the anti-resonant shunt circuit of Fig. 50 (b) and its analogue (a), since these are the duals of the circuits just considered, and the term $X$, above, merely represents 1/reactance instead of reactance.

This theorem is extremely useful, for in conjunction with the theorem concerning the transfer of a carrier wave to zero frequency (see Sec. 16), it greatly simplifies some circuit response calculations, as will be seen in the next chapter (Sec. 41).

## REFERENCES


1. L. C. JACKSON, "Wave Filters," Methuen Monograph, Methuen & Co., London, 1944.
2. F. E. TERMAN, "Radio Engineer's Handbook," Section 3, McGraw-Hill, New York, 1943 (or any standard book on Wave Filters).
3. H. WHEELER and F. MURNAGHAN, "Theory of Wave Filters containing a Finite Number of Sections," *Phil. Mag.* (7), 6, 1928, p. 146.
4. D. G. TUCKER, "Insertion Loss of (Single Section) Filters," *Wireless Engineer*, Vol. 22, Feb. 1945, p. 62.
5. C. W. MILLER, "Single Section *m*-Derived Filters," *Wireless Engineer*, Vol. 21, Jan. 1944, p. 4.
6. H. A. WHEELER, "Wideband Amplifiers for Television," *Proc. I.R.E.*, Vol. 27, July 1939, p. 429.
7. A. V. BEDFORD and G. L. FREDENDALL, "Transient Response of Multi-Stage Video-Frequency Amplifiers," *Proc. I.R.E.*, Vol. 27, April 1939, p. 277.
8. H. E. KALLMAN, R. E. SPENCER, and C. P. SINGER, "Transient Response," *Proc. I.R.E.*, Vol. 33, No. 3, March 1945.
9. A. A. BARCO, "Measurement of Phase-Shift in Television Amplifiers," *R.C.A. Review*, III, No. 4, April 1939.
10. T. E. SHEA, "Transmission Networks and Wave Filters," Van Nostrand, New York, 1929.
11. S. P. MEADE, "Phase Distortion and Phase Distortion Correction," *B.S.T.J.*, Vol. 7, April 1928, p. 195.
12. M. F. GARDNER and J. L. BARNES, "Transients in Linear Systems," Vol. 1, John Wiley (Chapman & Hall), 1942, p. 46.
13. R. M. FOSTER, "A Reactance Theorem," *B.S.T.J.*, Vol. 3, April 1924, p. 259.

14. E. Guillemin, "Communication Networks," Vol. II., John Wiley, N.Y.
15. E. K. Sandeman, "Generalised Characteristics of Linear Networks," *Wireless Engineer*, Vol. 13, Dec. 1936, p. 637.
16. N. I. Korman, "A Network Theorem," *Proc. I.R.E.*, Nov. 1944, Vol. 32, No. 11, p. 710.
17. V. D. Landon, "The Band-pass, Low-pass Analogy," *Proc. I.R.E.*, Dec. 1936, Vol. 24, No. 12, p. 1582.

The following references are also of interest:

(A) K. S. Johnson and T. E. Shea, "Mutual Inductance in Wave Filters," *B.S.T.J.*, Vol. 4, Jan. 1925, p. 52.
(B) G. A. Campbell, "Physical Theory of the Electric Wave Filter," *B.S.T.J.*, Vol. 1, Nov. 1922, p. 1.
(C) R. M. Foster, "Theorems Regarding the Driving-point Impedance of Two-mesh Circuits," *B.S.T.J.*, Oct. 1924.
(D) S. Butterworth, "On the Theory of Filter Amplifiers," *Wireless Engineer*, Vol. 7, Oct. 1930, p. 536.
(E) W. Cauer, "Topologische Dualitätsätze und Reziprozitätstheoreme der Schaltungstheorie," Zeit. für Angewandte Math. und Mech., Vol. 14, 1934, p. 349 (for original publication of the dual circuit construction).

## CHAPTER 4

# The Transient Response of Networks

### 34. The transmitting qualities of a channel

The various component stages which comprise a communications channel may be of two types:

(1) *Bilateral* types, which may transmit energy in either direction between the input and output terminals. For example, filter structures, lines, and attenuators have this property.

(2) *Unilateral* types, which transmit energy in one direction only, such as thermionic amplifiers and relays. With these the term "transmitted" energy is a misnomer, since the energy is supplied by a local source of power and is merely controlled by the unilateral device (e.g. the H.T. battery in an amplifier). Unilateral devices may further be divided into two classes: those which change the circuit configuration on receipt of a stimulus, such as telephone relays, and those which do not, such as the conventional amplifier. The former may change the number of meshes or branches in a circuit, and such methods of communicating a message are used only in the cases of comparatively simple signals and not for communicating the (ideally) exact waveform of a speech or vision signal. Other devices may vary the *condition* of a circuit without changing its configuration. For example, carbon and condenser microphones change respectively the resistance and capacity in a circuit.

When a signal is transmitted through the stages of a practical communications channel it must undergo some kind of distortion, so that the waveform appearing on the output terminals differs, in general, from that across the input terminal. If those stages are not considered at the moment which produce a change of circuit condition or configuration the distortion must arise from one of two causes:

(1) Non-linearity.

(2) Frequency selectivity.

Non-linearity in a channel means that the output signal level is not strictly proportional to the input level, and may be due to valve-characteristic curvature or to hysteresis in the iron cores of inductances and transformers, or other causes. Such non-linearity,

particularly in valves, must not be regarded as a perpetual nuisance, since on their non-linearity depends their usefulness as detectors, super-het mixers, etc. In straightforward amplifiers such non-linearity is, however, purely a source of distortion, since its effect is to produce new frequencies and to add them to the signal. Not only harmonic frequencies are produced, but when a complex signal is applied, sum and difference frequencies and harmonics of these appear.

If a curve is plotted relating the output/input signal level in a channel, then the method of assessing the distortion of an applied signal having a definite geometric waveform is obvious. But in the case of speech or similar signal consisting of an irregular waveform comprised partly of unrelated frequencies, the harmonic distortion may not readily be calculated, though methods of analysis have been produced for certain fairly simple cases.[1, 2]

The other form of signal distortion, namely that due to frequency selectivity in a channel, arises because the spectrum of the transmitted signal must be modified by such selectivity. Such selective effects can arise only in circuits having reactive elements (as well, possibly, as resistances) capable of storing energy, but since as a general rule all communications channels need to possess selectivity, distortion must always arise in them. Frequency selectivity is ultimately necessary for separating signals transmitted simultaneously, either over a wire or broadcast. In practical systems there is a definite maximum range of frequencies available into which a limited number of signals may be fitted, the carrier frequencies of the signals being spaced at intervals of frequency so that they and their associated side-band spectra may each be selected by appropriate band-pass filters. The selectivity of any filter must be such that the side-bands of the adjacent unwanted signals are substantially eliminated.

Clearly, the narrower the side-band spectra of each adjacent channel, the closer the carrier frequencies may be placed, and therefore the more channels may be used, requiring narrower band-width filters to obtain the required selectivity. It is well known that the less the band-width of a channel the poorer will be the quality of the signal received through it, but some quantitative assessment of this deterioration is needed. Merely to refer to the "band-width" of a channel is insufficient, since the exact form of the characteristics depends on the type of selective circuits used; furthermore, the expression "band-width" in this way is really meaningless, since the

sides of a practical filter characteristic are not infinitely steep and the band-width must be measured at some arbitrary level of attenuation.

By way of example of the typical forms of signal distortion that arise in a communications channel Fig. 51 illustrates the response of a common type of vision-frequency amplifier, as used in television. The signal $E$, applied to the input terminals $ab$, has a step waveform and the curves show the waveforms of the signals, $e_N$, on the output terminals, $cd$, for different numbers of valve stages $N$. For clarity,

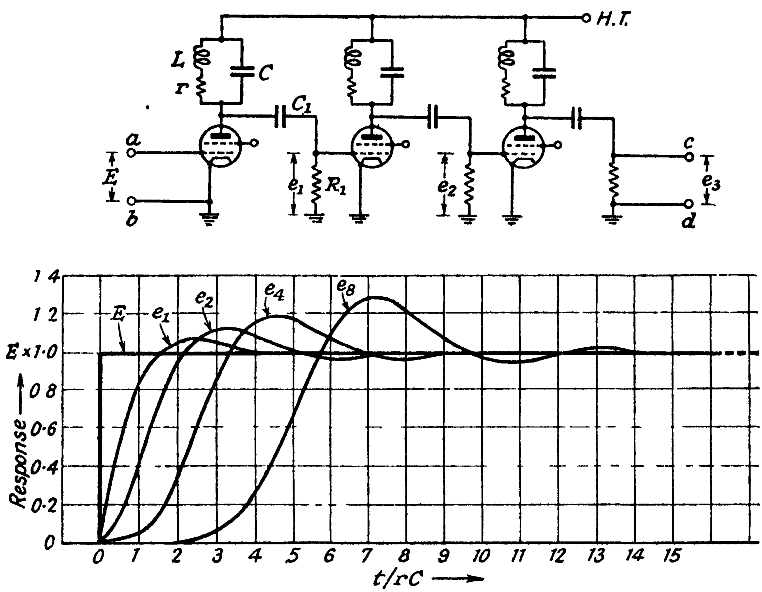



Fig. 51.—Response of an $N$-stage Amplifier to a Step-wave ($N$=1, 2, 4, 8 stages). $L/Cr^2=Q^2=0{\cdot}5$.

all these responses have been reduced to the same magnitude. These waveforms are distorted from the ideal step wave in three pronounced ways:

(1) The response takes a finite time in which to build up, and this time increases with $N$, the number of stages. This "build-up time" is usually quoted as the time taken to rise from 10 per cent. to 90 per cent. of full steady amplitude.

(2) There is an apparent time delay between the application of the step-wave and the output signal response, which also

increases with $N$. This "delay time" is usually measured to the half-amplitude point on the response waveform.

(3) The response overswings its steady maximum value by a certain percentage.

Of these, (1) and (2) are particularly important, and are characteristic of all types of channel—vision, speech, or telegraphy. In the case of speech* the signals may be considered to consist of a series of abrupt disturbances similar to the step wave, and the build-up time will have an effect on the intelligibility of the signal. The poorer the build-up time of a channel, the less capable is it of responding to rapid changes in the applied signal waveform and the poorer will be the definition of the received signal. The definition is also affected by the degree of overswing in the response waveform, and the magnitude, periodic time and decrement of this overswing depend entirely on the particular type of structure of the selective circuits used. In our present example the circuits are of an extremely simple form, being asymmetrically damped tuned circuits, and the overswing is decided by their Q value (a Q of $1/\sqrt{2}$ has been chosen for the illustration).

It is sometimes more convenient to consider the definition of a channel from its steady-state characteristics, particularly in the case of sound channels, since such characteristics tell just as much about the distortion of a signal as does the step-wave response. The two forms of response specification are closely linked by Fourier's theorem, which will be considered in further detail in this chapter. Fig. 52 shows the frequency characteristics of one stage of the amplifier we have exampled (Fig. 51). The corresponding step-wave response, of a single stage, is the curve $e_1$ on Fig. 51; these two responses, the steady-state and the transient, both determine the behaviour of this particular type of stage under all signal conditions, assuming that the valve is linear.

Two such stages give the step-wave response $e_2$, which has a poorer build-up time, and they also have a reduced band-width selectivity characteristic, however this band-width be measured. Thus in Fig. 52 the frequency component $\omega = 2{\cdot}5/rC$ is reduced to half-amplitude, so that with two similar stages this component will be reduced to a quarter-amplitude, with three stages an eighth-amplitude and so on. The build-up time gets progressively worse as the channel band-width reduces, a general rule attributed to

* Consult reference 6 for the spectra and composition of sound signals.

Kupfmuller.[3] (The foregoing is merely an example and not a proof.) Again, as $N$, the number of stages, is increased the phase-shift through the channel increases linearly with $N$, which causes the time delay of the step-wave response to increase, as will be seen later in further detail.

Both the steady-state response characteristics, Fig. 52, and the step-wave response curves, Fig. 51, have been drawn as "universal"

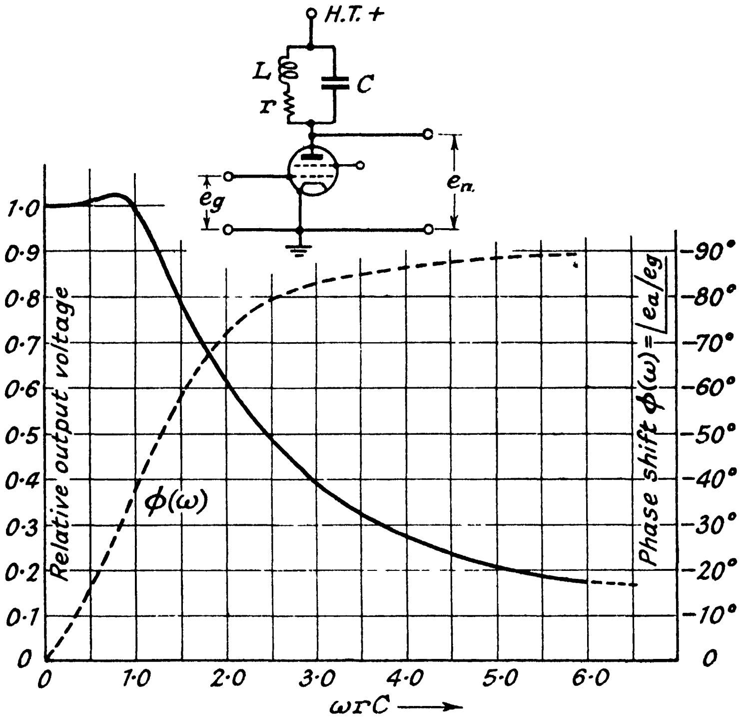



Fig. 52. The Steady-state Frequency Characteristics of an Amplifier Stage. $L/Cr^2 = Q^2 = 0{\cdot}5$.

curves against a scale of relative frequency $\omega rC$ and relative time $t/rC$, $rC$ being the time-constant of the selective circuit used. These characteristics then remain the same *shape* but vary in their scales if $rC$ be changed but if the Q value be kept constant. This may be done by keeping $r$ constant but varying $L$ and $C$ in the same ratio, in which case the waveform of the channel response will be maintained as regards shape but will change in scale, thereby transmitting the same information, but taking different times to do so; reducing

$L$ and $C$ sharpens up the step-wave response (on an absolute time scale) and widens the selective characteristic "band-width," thus improving the definition of the channel. This principle is not so obvious in cases of more complex selective circuits, containing a number of meshes, but it is applicable to every mesh in the circuit.* There are of course practical difficulties involved in certain types of communication channel by taking too long a time to transmit a given piece of "information," i.e. a waveform shape; for example, it would be objectionable in television, though less serious in facimile transmission. From the arguments set out here the general rule may be formulated that the total amount of information which may be sent over a given communication channel is proportional to product of the channel band-width and the time for which it is used.[4]

Although complete steady-state characteristics are more commonly used than transient response characteristics for specification of the quality of a channel, it must be appreciated that all communication signals must be transient in form, since the idea of transmitting information involves change of some kind.[5] A continuous sine wave cannot, by itself, convey a message, but it must be varied in some way. The response to transients may be calculated from the steady-state characteristics, though in many cases this may be impossible or extremely difficult, mathematically, but frequently it is sufficient to assess the transient behaviour in a qualitative way, which may be easier. The direct calculation of transient response may be performed by a solution of the differential equations of the circuits, for which purpose various operational means have been devised. Once the differential equations have been formulated the problem ceases to be a physical one and becomes a mathematical exercise and only too often, unfortunately, a difficult one.

## 35. Transient response from steady-state characteristics

The response of a network to a transient may be determined from a knowledge of the complete steady-state characteristics, if the magnitude and phase-shift distortion of every (Fourier) sinusoidal component in the applied transient, imposed by the network's

* It was pointed out in Chapter 1 that the waveform *shapes* of the normal modes of oscillation of a circuit depend only on the Q values of the meshes (see Fig. 8).

characteristics, be taken into account.[21] The calculation requires three distinct processes:

(1) Analysis of the applied transient into its Fourier spectrum.
(2) Modification of this spectrum according to the characteristics of the network.
(3) Summation of all the component sinusoidal waves in this modified spectrum.

The problem may arise in two ways. First, as a purely theoretical calculation, from a knowledge of the analytical functions representing the transient and the network's characteristics; secondly, as an approximation from practical measured characteristics or from numerical calculation of the characteristics at specified "spot" frequencies. Both problems involve the same principles.

Suppose a periodically repeated transient wave, of a known analytical form, be applied to the terminals of a network whose transfer characteristics are also known as analytical expressions:

$$\left.\begin{array}{l}\text{Let the applied wave be } i=f(t)\\ \text{and the network transfer characteristics } Z(\omega),\phi(\omega)\end{array}\right\} \quad . \quad (178)$$

Then the applied wave may be analysed into its Fourier harmonic components. If this operation is not possible or proves difficult, then the wave may be split up into a number of simpler waveforms, each of which may be analysed into its own series of components (for instance, the applied wave may be considered to consist of a number of step waves or rectangular pulse waves added together, as shown later in Sec. 42). The network's responses to each of these simpler waves may then be calculated and the results added, by the superposition principle, to give the resultant response. Suppose $f(t)$ in this instance is analysable and that its Fourier series of harmonic components is given by:

$$i=f(t)=\sum_{n=0}^{\infty}a_n \cos \omega_n t+b_n \sin \omega_n t \quad . \quad . \quad (179)$$

where $a_n$ and $b_n$ are the amplitudes of the $n^{th}$ cosine and sine harmonics respectively. The response wave of the network will have a similar spectrum of harmonics, but modified in amplitude and phase.

The network's transfer impedance at these harmonic frequencies will be given by:

$$\left.\begin{array}{ll}\text{Modulus:} & |Z(\omega_n)|\\ \text{Phase shift:} & \phi(\omega_n)\end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (180)$$

Thus the spectrum of the output response wave will be:

Response wave

$$=\sum_{n=0}^{\infty}|Z(\omega_n)|\,.\,\{a_n\cos[\omega_n t-\phi(\omega_n)]+b_n\sin[\omega_n t-\phi(\omega_n)]\}\quad . \quad (181)$$

which may be calculated in any practical case from the "spot"-frequency readings of $|Z(\omega_n)|$ and $\phi(\omega_n)$ whether these be calculated or measured.

The procedure in the case of a single transient [7] is very similar, but requires the transient to be analysed into its continuous spectrum as a Fourier integral. Suppose the applied transient to be of such a form that this spectrum may be calculated:

$$f(t)=\int_{-\infty}^{+\infty}a(\omega)\cos\omega t+b(\omega)\sin\omega t.d\omega \qquad \text{(see equation 114)}$$

This represents the sum of the continuous spectrum of cosine and sine component waves which comprise the transient. Now, the network transfer impedance, being in an analytical form, is known for every conceivable frequency, and so the response of the network to each component in the above spectrum is calculable. The spectrum of the output response transient will thus be given by:

Response transient

$$=\int_{-\infty}^{+\infty}|Z(\omega)|\,.\,\{a(\omega)\cos[\omega t-\phi(\omega)]+b(\omega)\sin[\omega t-\phi(\omega)]\}d\omega\quad . \quad (182)$$

This spectrum may be written in terms of pure cosine and sine components, by expanding 182:

Response transient

$$=\int_{-\infty}^{+\infty}|Z(\omega)|\,[a(\omega)\,.\cos\phi(\omega)-b(\omega)\sin\phi(\omega)]\cos\omega t.d\omega$$

$$+\int_{-\infty}^{+\infty}|Z(\omega)|\,.\,[b(\omega)\cos\phi(\omega)+a(\omega)\sin\phi(\omega)]\sin\omega t.d\omega\quad . \quad (183)$$

Thus the response may be expressed as a function of frequency, which is not suitable for many practical purposes since it is usually required as a function of time. This involves the evaluation of the integrals (183), that is to say the infinite spectra of sinusoidal components must be added together to obtain an expression for the resultant response waveform. Immediately a difficulty is presented in very many examples, since comparatively few Fourier integrals are soluble. A list of solutions of great practical use has been

given by G. A. Campbell,[8] but the method of attack, using steady-state theory, is unfortunately limited in this manner. However, the ideas behind the method are of great importance.

Campbell's list of integrals, together with many standard publications on Fourier analysis, uses the complex exponential notation. We have shown in Sec. 20 how this and trigonometrical notations are linked, but while the great neatness of the complex exponential notation is not to be underestimated, it may be that the trigonometrical form has certain advantages in this present instance, in that it makes more apparent the symmetry relations between the components of spectra and of network characteristics (i.e. between odd and even wave components, real and imaginary parts of characteristics, etc.). The following examples should make this clearer.

EXAMPLE 1. *Response of a Network to a Symmetrical Driving Waveform*

Let us consider the voltage response of a network of known* transfer impedance $Z(\omega),\phi(\omega)$, to an injected current of "rectangular" waveform. Such a waveform is illustrated in Fig. 53 (a) (and has been plotted accurately in Fig. 29 on the diagrams of pulse-type waveforms), and it is symmetrical about the axis $t=0$. This waveform is commonly used in calculations of the transient response of networks, and it will be well to evaluate its spectrum. The waveform $f_c(t)$ may be defined thus:

$$f_c(t)=I, \text{ for } -\frac{T_1}{2}<t<+\frac{T_1}{2}$$
$$=0, \text{ for } t \text{ outside these limits} \quad . \quad . \quad . \quad . \quad . \quad (184)$$

The spectrum of such a symmetrical wave is given by the Fourier integral (113), repeated here:

$$a(\omega)=\frac{1}{\pi}\int_{-\infty}^{+\infty} f_c(t)\cos\omega t.dt \quad . \quad . \quad . \quad . \quad [(113)]$$

Substituting for $f_c(t)$, from 184:

$$a(\omega)=\frac{1}{\pi}\int_{-T_1/_2}^{+T_1/_2} I\cos\omega t.dt=\frac{2I}{\pi\omega}\sin\frac{\omega T_1}{2} \quad . \quad . \quad . \quad (185)$$

which represents the amplitudes of a continuous spectrum of *cosine* terms (Fig. 53 (b)), since $f_c(t)$ is a symmetrical or even function.

* For purposes of illustration the network has been assumed to consist of a shunted resistance and condenser in the diagrams of Fig. 53.

Then $f_c(t)$ may be written as the sum of the terms of this infinite spectrum. Owing to the symmetry we may use the $\frac{1}{2}$ range integral:—

$$f_c(t)=\int_0 a(\omega)\cos\omega t d\omega$$

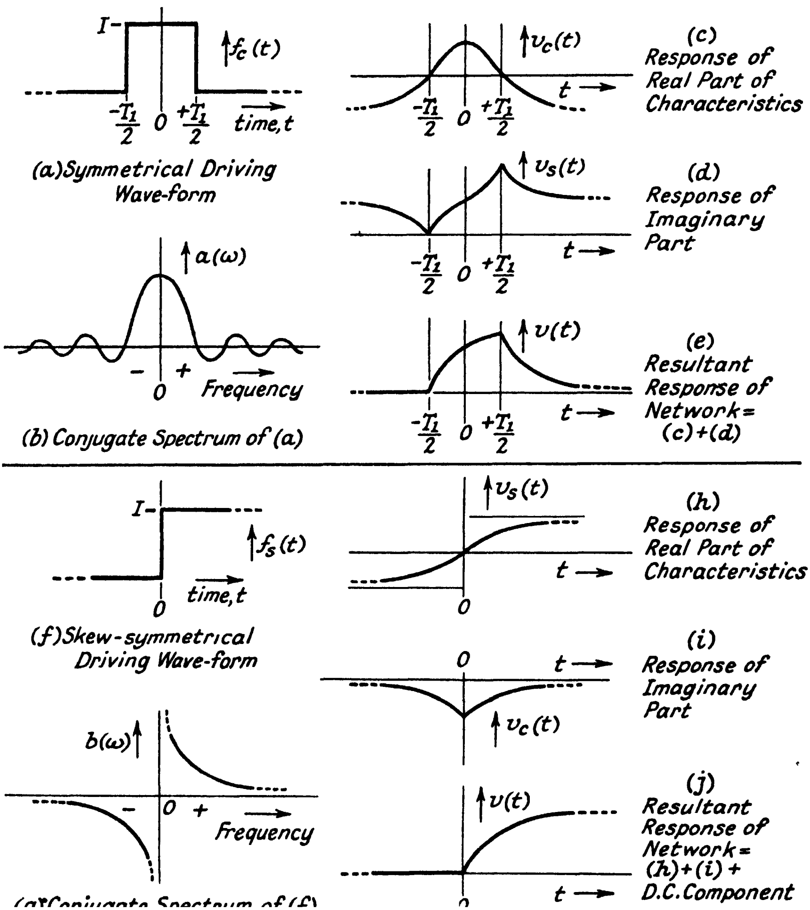



Fig. 53.—Response of a Network to a Symmetrical and a Skew-symmetrical Driving Waveform.

If this signal be applied to the network, the corresponding response voltage will be:

$$v(t)=\int_0^\infty |Z(\omega)\,.\,a(\omega)|\,.\cos[\omega t-\phi(\omega)].d\omega$$

$$v(t)=\int_0^\infty [|Z(\omega)\,.\,a(\omega)|\cos\phi(\omega)]\cos\omega t$$
$$+[|Z(\omega)\,.\,a(\omega)|\sin\phi(\omega)]\sin\omega t.d\omega.$$
$$v(t)=v_c(t)+v_s(t)\quad .\ .\ .\ .\ .\ .\ .\ .\quad (186)$$

the two components $v_c(t)$ and $v_s(t)$ here representing the responses of the real and imaginary parts of the network characteristics respectively (see Sec. 26, equation 131). Thus the response of the real characteristic, $v_c(t)$, will consist of *cosine* terms and will still be a symmetrical wave. The response of the imaginary characteristic $v_s(t)$ will consist of *sine* terms and must therefore be an odd function, or skew-symmetrical wave. These separate responses are illustrated in Fig. 53 (c) and (d).

But in a physical network these two parts cannot be separated and the two responses, even and odd, of equation 186 must be added together as shown. But since the applied current wave $i=f_c(t)$ is of zero amplitude before the time $t=-T_1/2$, there cannot be any response from the network before this time; thus:

$$v(t)=0 \text{ for } t<-T_1/2\quad .\ .\ .\ .\ .\quad (187)$$

as shown in Fig. 53 (e), and the two components, symmetrical and skew-symmetrical, must cancel before this time. This necessitates an important relation between the real and imaginary characteristics of a physical network, which will be investigated further in Sec. 43.

EXAMPLE 2. *Response of a Network to a Skew-symmetrical Driving Waveform*

If the injected current waveform be skew-symmetrical about $t=0$, for example as in the step wave in Fig. 53 (f), its spectrum must consist of *sine* components only, since this waveform function, $f_s(t)$, is odd.

The waveform $f_s(t)$ may be defined thus:

$$\left.\begin{array}{l} f_s(t)=0 \text{ for } t<0 \\ f_s(t)=I \text{ for } t>0 \end{array}\right\}\quad .\ .\ .\ .\ .\ .\ .\quad (188)$$

and the wave will have an average or D.C. value of $I/2$. This waveform is also commonly used in transient calculations and it forms the basis of the Heaviside operational method of analysis,[9] for which purpose it is called the Unit Function. Its spectrum may be calculated by the methods outlined in Secs. 20 and 21, though direct substitution of the conditions 188 into the Fourier integral

113 leads to an impossible integration. It may, however, be evaluated as follows, with reference to Sec. 20.

Consider the periodic square wave of period $T_0$ and amplitude $I$ as in Fig. 25 (a), with the time origin at point (2). This wave may be defined as (per cycle):

$$\left.\begin{array}{l} f_s(t)=0 \text{ for } t \text{ between } -T_0/2 \text{ and } 0 \\ f_s(t)=I \text{ for } t \text{ between } 0 \text{ and } +T_0/2 \end{array}\right\} \quad . \quad . \quad (189)$$

Then the amplitude of the $n^{th}$ harmonic will be given by equation 93:

$$b_n=\frac{2}{T_0}\int_0^{T_0/2} I \sin\frac{2\pi nt}{T_0}\,dt \quad . \quad . \quad . \quad . \quad . \quad (190)$$

giving
$$b_n=\frac{I}{\pi n}(1-\cos\pi n) \quad . \quad . \quad . \quad . \quad . \quad . \quad (191)$$

But $\cos \pi n=-1$ or $+1$ according as $n$ is odd or even. The amplitudes of the spectrum of odd *sine* harmonics for this square wave are then:

$$b_n=\left(\frac{2I}{\pi n}\right)_{n\ odd} \quad . \quad . \quad . \quad . \quad . \quad . \quad (192)$$

Now let the period $T_0$ become indefinitely large so that the square wave becomes the step wave of Fig. 53 (f). The fundamental frequency becomes indefinitely small and the harmonics crowd closely together. Let the odd harmonic spacing be $\delta\omega$ (i.e. equal to twice fundamental frequency, since there are only alternate harmonics) then $n\delta\omega/2$ becomes the general frequency $\omega$. The spectrum becomes a continuous function $b(\omega)$ so that equation 192 may be rewritten:

$$b_n/\delta\omega \rightarrow b(\omega)=\frac{I}{\pi\omega} \quad . \quad . \quad . \quad . \quad . \quad (193)$$

These are the amplitudes of *sine* terms; adding all these together, from zero to infinite frequency, results in the step wave, represented in this way by the Fourier integral:

$$f_s(t)=\int_0^\infty b(\omega)\sin\omega t\,d\omega=\frac{I}{\pi}\int_0^\infty\frac{\sin\omega t}{\omega}d\omega \quad . \quad . \quad (194)$$

to which must be added the D.C. component $I/2$, since the wave is unidirectional. This spectrum $b(\omega)$ is illustrated in the conjugate form in Fig. 53 (g), and is skew-symmetrical.

Now if this current be injected into the network, the terms in the spectrum will be modified in amplitude and phase by the characteristics of the network, $|Z(\omega)|$ and $\phi(\omega)$. The voltage response wave will then be:

$$v(t)=\int_0^\infty |Z(\omega) \,.\, b(\omega)| \,.\, \sin [\omega t-\phi(\omega)]d\omega$$

$$=\int_0^\infty [|Z(\omega) \,.\, b(\omega)| \cos \phi(\omega)] \sin \omega t$$

$$-[|Z(\omega) \,.\, b(\omega)| \sin \phi(\omega)] \cos \omega t.d\omega$$

$$=v_s(t)+v_c(t) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (195)$$

to which may be added the D.C. component $(Z(0) \,.\, I/2)$.

The first component, $v_s(t)$, is the response of the real part of the network characteristic and is an odd function, therefore representing a skew-symmetrical wave, while the second component $v_c(t)$ is even and therefore a symmetrical wave (see Fig. 53 (h) and (i)). The two parts are inseparable in a physical network, and so their sum $v(t)$ must be of zero amplitude before $t=0$, the time at which the step-wave current is injected.

Although we have referred here specifically to an injected current and a response voltage, this has been for illustration only, and the conclusions may be applied to any driving wave and its corresponding response wave provided that the network characteristics are defined to suit (see Sec. 24, Chapter 3).

## 36. Example of Fourier integral analysis of circuit response

As an example of the calculation of the transient response of a circuit by Fourier integral analysis, let us consider the very simple case of a resistance-loaded tetrode amplifier stage in which the valve anode to earth shunt capacity is appreciable; such a case may occur in television. It is not suggested that this method is the preferable one in such a case, but it is given here as an example of the Fourier integral application, outlined in Sec. 35.

If a step-wave voltage, $e$, be applied to the valve grid (see Fig. 54), the anode current waveform will be of the same shape, assuming perfect decoupling of the screen and cathode. Then a steady current $I$ will suddenly be passed through the anode load, starting at the instant $t=0$; we require to know the anode voltage response.

The spectrum of such a wave has already been calculated (equation 193), the infinitesimal components being *sine* waves since the step

wave is an odd function. A component of frequency $\omega$ will have the amplitude:

$$d\omega . b(\omega) = \frac{I d\omega}{\pi\omega} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (196)$$

The step wave and its spectrum are illustrated by Fig. 53 (f) and (g).

Now the impedance of the anode load, at frequency $\omega$, is:

$$Z(j\omega) = \frac{R(1 - j\omega CR)}{(1 + \omega^2 C^2 R^2)} \quad . \quad . \quad . \quad . \quad . \quad (197)$$

The component of the response voltage at frequency $\omega$ will then have (from 196 and 197) the value:

$$d\omega . b(\omega) \,.\, Z(j\omega) = \frac{IR(1 - j\omega CR)}{\pi\omega(1 + \omega^2 C^2 R^2)} d\omega \quad . \quad . \quad . \quad (198)$$

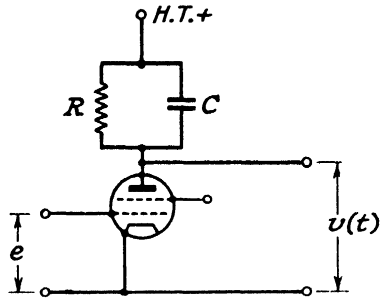



Fig. 54.—Tetrode Valve with Resistance-capacity Load.

This expression (198) gives the spectrum $b(\omega) \,.\, Z(j\omega)$ of *sine* components of the response voltage wave. The waveform of this response $v(t)$ is then given by the sum of all the components in this spectrum from zero to infinite frequency, that is:

$$v(t) = \frac{IR}{\pi} \int_0^\infty \left[ \frac{\sin \omega t}{(1 + \omega^2 C^2 R^2)\omega} - \frac{j\omega CR \,.\, \sin \omega t}{(1 + \omega^2 C^2 R^2)\omega} \right] d\omega$$

$$= v_s(t) + v_c(t) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (199)$$

illustrated by Fig. 53 (h), (i), (j).

The integral has been written in this way to show the responses of the real and imaginary parts of the circuit separately. As we saw in the last section, these two responses are inseparable in practice and hence must be of the same form but opposed in

symmetry, so that they add up to zero before time $t=0$. We need therefore calculate either $v_s(t)$ or $v_c(t)$, but not both. Let us take $v_c(t)$:

$$v_c(t)=-\frac{IR}{\pi}\int_0^\infty \frac{jCR\,.\,\sin\omega t}{(1+\omega^2C^2R^2)}d\omega \quad . \quad . \quad . \quad (200)$$

This is a standard Fourier integral, and its value as an explicit function of time may be found in a list of such integrals.*

$$v_c(t)=-\frac{RI}{2}\,.\,\epsilon^{-t\,RC} \quad . \quad . \quad . \quad . \quad . \quad (201)$$

Similarly the response $v_s(t)$ of the real part of the characteristic must be of the same form for $t>0$, but equal and opposite for $t<0$ (see Fig. 53 (h) and (i)) so that the total response will be (including the D.C. component $RI/2$):

$$\left.\begin{aligned} v(t)&=RI(1-\epsilon^{-t/RC}) \text{ for } t>0 \\ &=0 \text{ for } t<0 \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (202)$$

(Fig. 53 (j)), a result which the reader will agree could have been obtained much more quickly by the solution of an elementary differential equation. However, the example illustrates the three essential processes for transient calculations on a steady-state basis: the processes of Fourier analysis, modification of the resulting spectrum by the network, and Fourier synthesis or integration of the resulting wave. The method is not workable in many practical cases with more complicated circuits, since only a limited number of Fourier integrals have been listed. The method also brings out the essential relation between the real and imaginary parts of the characteristics of a practical network.

The Fourier integral approach will be seen to be of more practical use for making estimations of the behaviour of circuits when we come to deal with idealisation in circuit analysis.

* In lists of these integrals, such as G. A. Campbell's[8] the variable $p(=j\omega)$ is used. If we use instead $p=j\Omega$, where $\Omega=\omega CR$, our "relative frequency" of Chapter 3, the Fourier integral (200) could be rewritten in the complex notation form (see Sec. 20) as $v_c(t)=-\frac{IR}{\pi}\int_{-\infty}^{-\infty}\frac{1}{(1-p^2)}\,.\,\epsilon^{-j\Omega t/RC}d\Omega$ which is the Fourier integral of $-IR\,.\,1/(1-p^2)$ and is given by Campbell's list number 444.

## 37. Practical transient response computation—the use of Fourier series approximation

Although the applications of Fourier integral analysis may be very limited when used in an analytical way, owing to the great difficulty of solving the resulting integrals, the ideas behind the method may be put to use for practical computation of transient response which can be extremely useful for finding an approximate answer. No analysis need be necessary, but the solution may be obtained by direct arithmetical addition of the various sinusoidal components forming the applied driving wave, after modifying their amplitudes and phases according to the characteristics of the network. These characteristics need not then be known as analytical expressions, but indeed may be measured curves. If the applied driving wave is transient in form the addition of an infinite number of sinusoidal components is required for an exact solution, but this can be avoided [10, 11] by using only certain chosen components spaced at equal frequency intervals: this means representing the transient by a Fourier *series*, as though the transient were repeated at regular periods of time.

In practical cases only one type of driving waveform need be considered—the symmetrical "square wave" of Fig. 53 (a)—since the transient response of a circuit to any other type of driving wave may be estimated from this. The amplitude spectrum of such a wave is given by equation 185; all the components, being cosine waves, have their peak values at $t=0$. Fig. 55 (a) shows a plot of this continuous spectrum, and (b) shows it represented by a series of harmonic terms spaced at intervals $\omega=2\pi/T_0$. The single square wave and repeated square wave (period $T_0$) are shown in Fig. 55 (c) and (d), corresponding to these two spectra. These harmonics may now be modified in amplitude and phase according to the network's characteristics, and the resulting series of waves added up to give the waveform of the network's response to the applied *repeated* wave.

The question arises of choosing a suitable harmonic spacing $2\pi/T_0$. The wider this spacing the less labour is required in the computation, but the shorter will be the repetition period $T_0$ and the less accurately will the eventual solution represent the correct transient solution. The criterion is the degree to which, in the resulting computed waveform, the transient response corresponding to one square wave has died away before the next square wave starts—the residual

voltage, $e_1$, in Fig. 56 (b) represents an error caused by choosing too small an initial repetition period $T_0$. Some error must always be

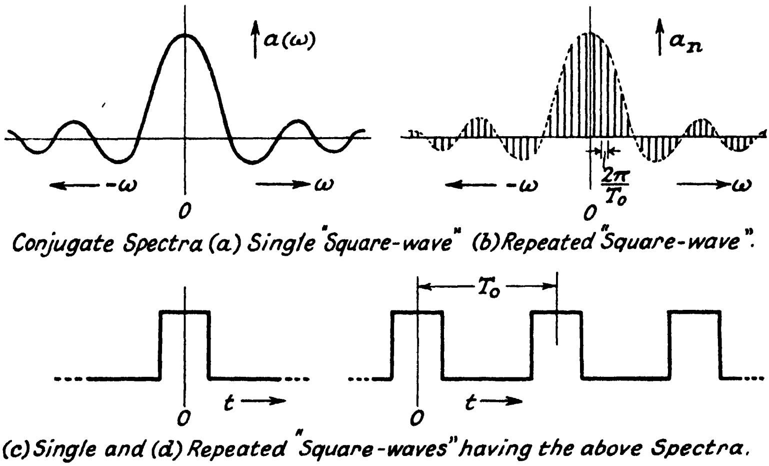



Fig. 55.—Spectra of Single and Periodic "Square-waves."

caused by this assumption of a periodic wave, and the effects of such error must be judged in any particular case. Although the response build-up curve (*ab* in Fig. 56 (c)) may last a finite time,

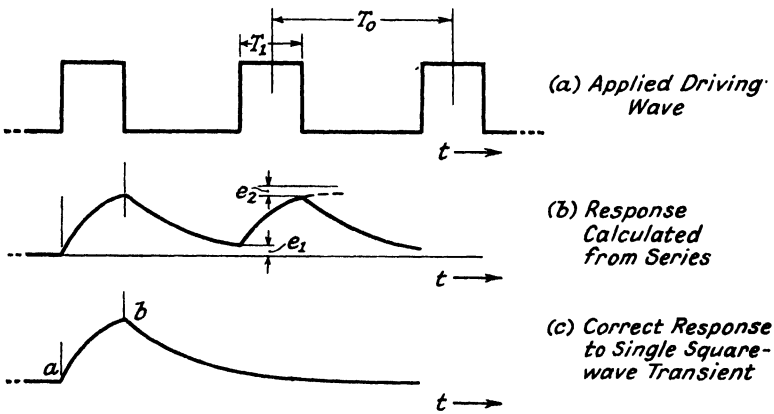



Fig. 56.—Error Due to Use of Fourier Series Approximation.

the decay time (after *b*) may be infinite with an applied signal of the form used here, in reactive circuits. This infinite decay time is only of theoretical interest, and is due to the fact that energy in

the reactances may take an indefinitely long time to dissipate, although in any practical case there must be some chosen level of residual voltage which may be considered negligibly small.

There is no exact way of arriving at a suitable value for $T_0$ before starting the computation, but frequently it is possible to make an estimate of the decay rate of the transient response curve from the known time-constants in the circuit, or alternatively from the steady-state characteristic curves. If these characteristics show sharp resonances, rapid changes in phase, or sharp rates of cut-off,

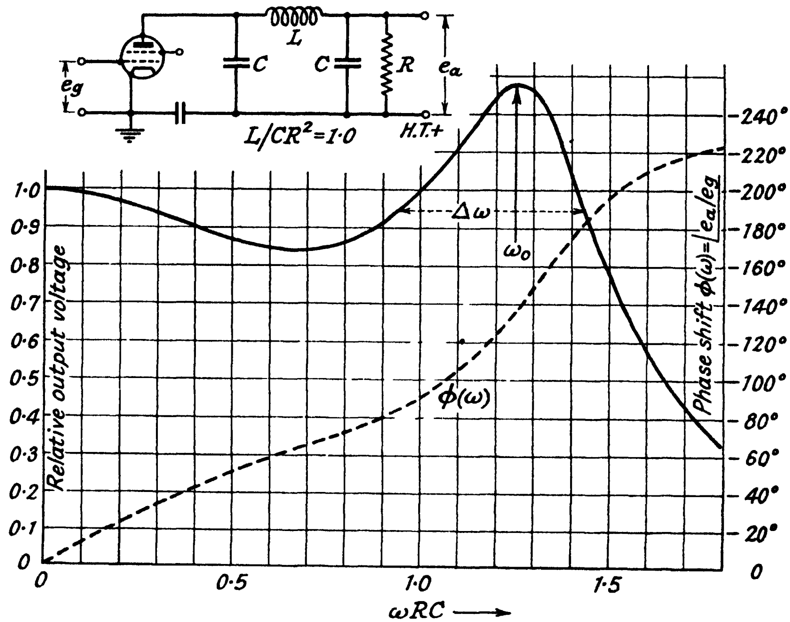



Fig. 57.—Characteristics of a "Video-frequency" Amplifier Showing Pronounced Peak.

then there will be a correspondingly large transient distortion. In such cases the repetition period $T_0$ must be chosen so that the harmonics of $2\pi/T_0$ cover these rapid changes in characteristic, thereby including their effects in the computation.

For example, the characteristics in Fig. 57 are those relating to a tetrode "video-frequency" amplifier loaded with a constant-$K$ filter section as shown in the circuit. The modulus of the characteristic has a peak, corresponding to a tuned mesh with a certain value of Q. This peak has not the exact form of a simple

anti-resonance curve, but in the region of anti-resonance it is approximately correct to regard it as being produced by an anti-resonant circuit and to estimate its effective Q from the band-width $\Delta\omega$ and tuning frequency $\omega_0$:

$$Q=\omega_0/\Delta\omega \quad \text{(Sec. 32, equation 169)}$$

where $\Delta\omega$ is measured at $1/\sqrt{2}$ of peak amplitude.

$$\left.\begin{aligned} \text{In the present example:}\quad \omega_0&=\frac{1\cdot25}{CR}\\ \Delta\omega&=\frac{0\cdot5}{CR}\\ \text{Then effective } Q&=2\cdot5 \end{aligned}\right\} \quad . \quad . \quad . \qquad (203)$$

The transient response of simple tuned circuits has been discussed in Chapter 1 (Sec. 5) and the relation between the Q and the ratio of decrement to natural frequency is expressed by equation 39. The waveforms of the responses are illustrated by the diagrams in Fig. 8, and from these an estimate may be made of the time required for the oscillations to decay to an inappreciable amplitude; this period of time may be regarded as suitable for the fundamental period, $T_0$, of our Fourier series approximation.

This is a particularly simple case; other examples of frequency characteristics may contain more than one sharp peak, but at least a rough estimate may be made in this way of the effective duration of the transient response, sufficiently accurate to provide a basis for the choice of the minimum repetition period $T_0$ of the repeated square wave used in the Fourier series computation.

It may be required to calculate the response of a network to a step wave instead of a single square wave, and this response may also be computed with reasonable accuracy. In this case, since we have already seen that a step wave may be regarded as a square wave of period tending to infinity (see Sec. 35), our Fourier series approximation will be sufficiently accurate provided that the duration $T_1$ (see Fig. 56) is great enough so that the computed transient response may reach its steady constant amplitude during this time $T_1$. The residual voltage $e_2$ is a measure of the error introduced by the use of too short a time $T_1$.

When making such approximate calculations the computer should first choose suitable values of $T_0$ and $T_1$ (it is advisable to let $T_0=2T_1$ so as to simplify the Fourier series) and then decide on suitable time intervals ($\delta t$) at which to calculate points on the

required transient response curve. Values of the fundamental component of period $T_0$ may be read off from trigonometrical tables at this interval ($\delta t$), adjusted in amplitude and phase according to the frequency characteristics of the network at this particular frequency, and then tabulated. Similarly the harmonics of the square wave may be dealt with. There is no advantage in considering harmonics of periodic time less than the interval ($\delta t$). Finally, the amplitudes of all these component response waves may be summed at the times $t=0$, $\delta t$, $2\delta t$, $3\delta t$, etc., giving the transient response.

The most laborious part of such a computation will be found to lie in the inclusion of the phase-shift characteristic, and so it is recommended as an alternative that the network's frequency characteristics be converted into real and imaginary parts before starting.

## 38. Conditions that a channel shall be distortionless

There are two essential conditions that a communications channel must satisfy in order that the waveforms of the input and output signals may be identical (apart from a probable difference of level). The channel must be linear and it must also be completely non-selective as regards frequency. The first requirement is simple enough, but the second needs some elaboration.

A given waveform may be analysed into a spectrum of sinusoidal terms, and that particular spectrum (of amplitude and phase angle components or alternatively sine and cosine components) is unique for that waveform. Hence if the amplitudes or phases of these components are disturbed by selective circuits, the waveform of the output signal will be modified. Let $|Z(\omega)|, \phi(\omega)$ be the transfer characteristics of a channel, having a signal $f(t)$ applied to its input terminals represented by a Fourier integral of cosine and sine terms:

$$f(t)=\int_{-\infty}^{+\infty} a(\omega)\cos\omega t+b(\omega)\sin\omega t\,.\,d\omega$$

the output signal $v(t)$ will then be, as in equation 182:

$$v(t)=\int_{-\infty}^{+\infty}|Z(\omega)|\ .\ \{a(\omega)\cos(\omega t-\phi(\omega))+b(\omega)\sin(\omega t-\phi(\omega))\}d\omega$$

If $f(t)$ and $v(t)$ are to be identical, the transfer characteristics must be of the form:

$$\left.\begin{aligned}|Z(\omega)|&=R\\ \phi(\omega)&=\omega t_1+n\pi\end{aligned}\right\}\quad\ldots\ldots\quad(204)$$

where $R$ and $t_1$ are constants and $n$ is an integer. Such perfect characteristics are not attainable in any channel containing reactance, but are ideals. If the spectrum of the transmitted signal occupies only a certain frequency band the above conditions need hold over that band only. These conditions are aimed at in most low-pass and band-pass filters. These ideal characteristics are illustrated by Fig. 58.

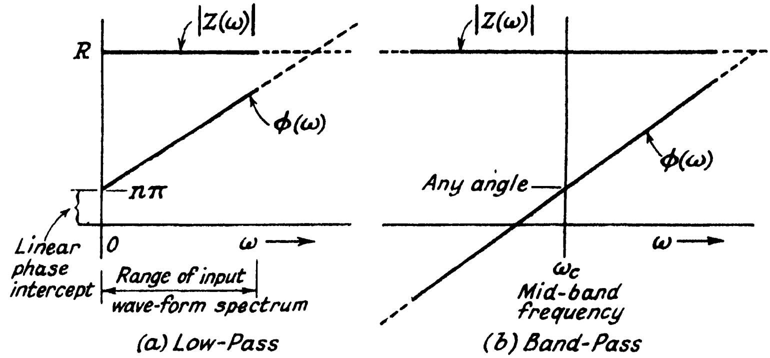



Fig. 58.—Ideal Channel Characteristics for Distortionless Transmission.

Substituting these conditions (204) into the equation for $v(t)$, the output wave, gives:

$$v(t)=\int_{-\infty}^{+\infty} R[a(\omega)\cos(\omega t-\omega t_1-n\pi)+b(\omega)\sin(\omega t-\omega t_1-n\pi)].d\omega$$

$$\text{or}\quad v(t)=R\int_{-\infty}^{+\infty} a(\omega)\cos(\omega\overline{t-t_1})+b(\omega)\sin(\omega\overline{t-t_1}).d\omega \quad . \quad . \quad . \quad (205)$$

which is identical with the input signal waveform except for a change in level $R$ and a time delay $t_1$. Thus we can see that a phase characteristic with a uniform slope $t_1=\phi(\omega)/\omega$ produces a delay in time of the entire transmitted wave, without waveform distortion. The phase shift at zero frequency, $n\pi$, is known as the *linear phase intercept*, and if $n$ be an odd number the wave is merely reversed in sign. If $n$, however, is not an integer the waveform will be distorted. For instance, if $n=\frac{1}{2}$ the cosine components become sine and vice versa, which may cause extreme modification of waveshape, as is illustrated by Fig. 59, which shows a "rectified cosine" pulse wave $f_c(t)$ and the corresponding waveform $f_s(t)$ having an identical amplitude spectrum but with every component shifted by $\pi/2$.

Thus in television amplifiers, radar pulse amplifiers and similar

"waveform transmitting channels" these ideals are aimed at—constant amplitude/frequency characteristics, a uniform phase slope, and a linear phase intercept an exact multiple of $\pi$.

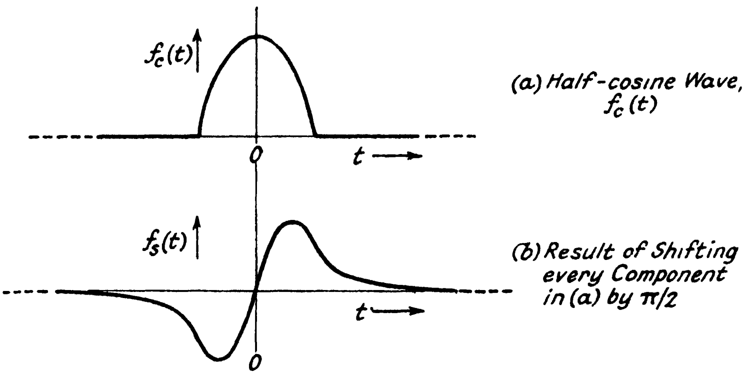



Fig. 59.—The Effect of a Linear Phase Intercept of $\pi/2$ on a Half-cosine Waveform.

## 39. Amplitude and phase distortion

The effect of departure from the ideal transmission characteristics discussed above is to produce transient distortion, and such distortion may be typical of either amplitude or phase errors or of the two in combination.

In a practical network the two parts of the characteristics, amplitude and phase shift, are inseparable and it is of purely theoretical interest to discuss the effects of each separately, but to do so will shed some light on the essential relations that must exist between these two parts.

We have already established the connection between the forms of the real and imaginary parts of the characteristics of a physical network (Sec. 35) and there must similarly be some kind of relation between the amplitude and phase shift; moreover, amplitude and phase shift are the more commonly used characteristics of a network and it is very convenient to be able to make some assessment of the signal distortion produced by their variation from the ideal linear forms.

If a network could be constructed to have a varying amplitude characteristic, $|Z(\omega)|$, but a perfect linear phase-shift characteristic, then any signal passed through it would suffer a symmetrical distortion of its waveform. For example, suppose a test signal of a symmetrical pulse form (Fig. 53 (a)) is applied to the network.

Such a pulse must contain *cosine* components only, and may be expressed as:

$$f_c(t)=\int_{-\infty}^{+\infty}a(\omega)\cos\omega t.d\omega$$

so that the response of our network will be the wave:

$$v_c(t)=\int_{-\infty}^{+\infty}a(\omega)\,.\,|Z(\omega)|\,.\cos\omega\overline{t-t_1}.d\omega \quad . \quad . \quad (206)$$

where $t_1$ is again the slope $\phi(\omega)/\omega$ of the assumed uniform phase-shift characteristic. Such a response wave also consists entirely of cosine terms and must therefore still be symmetrical. This means that the pulse must be distorted equally on either side of $t=t_1$ so that a response could be made to appear before the time of application of the leading edge of the input signal ($t=-T_1/2$). Furthermore, the trailing edge of the response, in any physical reactive network, must extend to infinity,* although its value after some time may be insignificantly small (but nevertheless finite), and so for perfect symmetry the leading edge would need an infinite time to build up. Therefore the phase-shift characteristic of a network must be of such a form as to introduce sufficient asymmetry into the response just to reduce this response to zero prior to the time of application of the leading edge of the input signal, $f_c(t)$.

Physical networks may be constructed having constant-amplitude characteristics but non-uniform phase-shift characteristics. Such networks introduce time delay in a transmitted signal and their phase characteristics may be made to assume a variety of forms. They are used [12] as "phase-correcting" circuits to lessen the distortion introduced by the non-uniform phase-shift characteristics of filters and other networks.

Let $\phi(\omega)$ be the non-uniform phase-shift characteristic and $R$ the constant-amplitude characteristic. If the input signal is the same pulse wave as before, $f_c(t)$, the response in this case will be:

$$v(t)=R\int_{-\infty}^{+\infty}a(\omega)\cos(\omega t-\phi(\omega)).d\omega$$

or, expanding:

$$v(t)=R\int_{-\infty}^{+\infty}[a(\omega)\,.\cos\phi(\omega)]\cos\omega t+[a(\omega)\sin\phi(\omega)]\sin\omega t.d\omega$$

$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (207)$$

* The energy stored in reactances can never be said to have dissipated completely at any given time, except when damping is *exactly* critical.

which consists of cosine as well as sine components. Thus a certain degree of asymmetry is introduced by a non-uniform phase-shift characteristic; such asymmetry does not imply the introduction of a finite response before the leading edge of the input waveform.

These characteristic forms of amplitude and phase-shift distortion are illustrated by Fig. 60.

The rates of build-up and of decay of the response signal depend on both the amplitude and phase-shift characteristics. These rates increase nearly proportionally to the band-width of the amplitude characteristic though the exact shape of the response curves will depend on the form of this characteristic. The phase-shift charac-

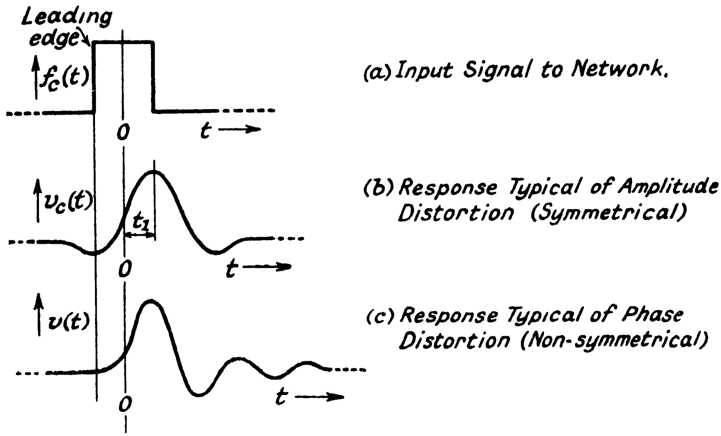



Fig. 60.—Illustrating Amplitude and Phase Distortion of a Rectangular Pulse Signal.

teristic has the effect of introducing some asymmetry in the response signal but cannot modify the energy in the signal; hence, with reference to Fig. 60, the areas of the responses with ordinates squared, before and after phase distortion, must be identical.

$$\int_{-\infty}^{+\infty}[R\,.\,f_c(t)]^2dt=\int_{-\infty}^{+\infty}[v(t)]^2dt \quad . \quad . \quad . \quad . \quad (208)$$

The general effect of a non-uniform phase-shift characteristic is then to remove part of the energy from the beginning of the signal element and replace it later in the signal, thereby producing asymmetry. This means that certain parts of the signal are delayed more than others, and such effects may only be produced by frequency discrimination in a linear network. Phase-shift distortion

can be very objectionable in systems such as television,[22] where faithfulness of reproduction of transient waveshape is important, since the delaying of certain portions of the signal more than others can give rise to a long "tailing" decay in the response and sometimes to overswings in this tail. The measurement [13, 15, 19] and correction [12, 14] of phase characteristics in such systems is therefore important.

## 40. Phase delay and envelope delay

We have seen in Sec. 38 (equation 205) that the time delay of any one sinusoidal component in a signal is equal to the ratio $\phi(\omega)/\omega$. If this is a constant for all frequencies the entire signal is delayed by this time, but is undistorted in waveform. If the phase charac-

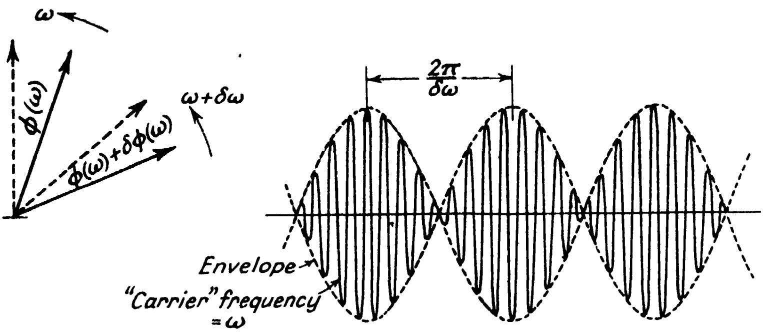


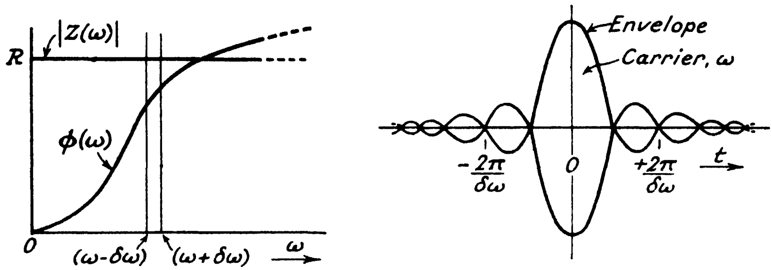

Fig. 61.—Envelope Delay—the Signal Corresponding to Two Adjacent Components of Frequency $\omega$ and $(\omega+\delta\omega)$.

teristic does not have a uniform slope, as is usual, the average slope will settle the delay of the bulk of the signal, but the distortion of the signal will bear little relation to the values of $\phi(\omega)/\omega$ for each component.

What tells us more about the signal shape distortion is the *relative* phase shift between neighbouring components, and this is perhaps best illustrated by taking two adjacent components of frequencies $\omega$ and $(\omega+\delta\omega)$ having phase shifts $\phi(\omega)$ and $\phi(\omega)+\delta\phi(\omega)$. The signal corresponding to these two components will be the "beat" wave shown in Fig. 61, having the envelope frequency $\delta\omega$. The two vectors shown there have angular velocities $\omega$ and $(\omega+\delta\omega)$ and come into phase at regular times giving the peaks of the resultant wave envelope. If the vectors be shifted in phase by $\phi(\omega)$ and $\phi(\omega)+\delta\phi(\omega)$ their relative shift is $\delta\phi(\omega)$ and

hence the time taken for the two vectors to "catch up" this angle will be $\delta\phi(\omega)/\delta\omega$, so that this extra time must elapse before the envelope now reaches its peak value. This time $\delta\phi(\omega)/\delta\omega$ is called the "envelope delay" time, and is equal to the slope of the tangent to the phase curve at the frequency $\omega$; it is in general different from the phase delay of the single component of frequency $\omega$.

Thus curves of envelope delay plotted against frequency are more indicative of transient distortion [17] than are curves of phase delay, $\phi(\omega)/\omega$. The contributions to the response signal of different parts of the energy spectrum is perhaps even more clearly seen by considering that part of the spectrum transmitted by a narrow element of the network characteristics (see Fig. 62 (a)) lying between frequency limits $(\omega-\delta\omega)$ and $(\omega+\delta\omega)$. It is assumed that the phase-




(a) Network Characteristics—Uniform Amplitude and Non-uniform Phase.

(b) Signal Element Corresponding to Frequency Group $(\omega \pm \delta\omega)$. (Envelope shown only.)

Fig. 62.—Envelope Delay—the Signal Element Corresponding to One Frequency Group.

shift characteristic is of uniform slope and the amplitude characteristic constant over this narrow range. Such a continuous spectrum may be called a *frequency group* and represents a pulse type of signal of the form* shown in Fig. 62 (b). This pulse of energy is a modulated carrier wave of frequency $\omega$ (of which the envelope only is shown in the Figure), and there will be a similar pulse element for every frequency group into which we divide the whole characteristic. The expression for the envelope is $(\sin \delta\omega \,.\, t)/\delta\omega \,.\, t$. The delay of the envelope of each such pulse will again be given by $\delta\phi(\omega)/\delta\omega$, if $\delta\omega$ is $<<\omega$, and so from this "envelope delay curve" may be seen the relation, in a general way,

* This signal is plotted accurately in Fig. 29—it may be derived from the shape of the spectrum of a rectangular pulse by using the reversible property of Fourier transforms.

between the non-uniformity of the phase characteristic and the build-up and decay of the signal response. If, at the same time, the amplitude characteristic of a network be non-uniform, this relation becomes extremely complex, since the frequency group of Fig. 62 is no longer of uniform amplitude and so the signal element will be distorted from its symmetrical form. The contributions to the entire response from each frequency group will then differ and simple curves of $\delta\phi(\omega)/\delta\omega$ are insufficient [16] for estimation of the complete signal distortion.

The frequency group which contributes mainly to the beginning of the build-up curve of a response signal, if the response be due to the application of a step-modulated carrier $\omega_0$ (see Fig. 64 (a)), will be that group which has the minimum time delay and will be located at the point corresponding to minimum $\delta\phi(\omega)/\delta\omega$. But the bulk of the energy in such a signal is centred around the carrier frequency $\omega_0$ and so the main part of the response signal is delayed by the envelope delay time at $\omega_0$. Therefore the very minimum build-up time of such a signal must be:

$$\left[\frac{\delta\phi(\omega)}{\delta\omega}\right]_{\omega_0} - \left[\frac{\delta\phi(\omega)}{\delta\omega}\right]_{MIN} \quad . \quad . \quad . \quad . \quad (209)$$

If the input signal is of such a form that it has a fairly uniform spectrum (for instance a narrow pulse) then the response signal must have an appreciable amplitude for a time lasting:

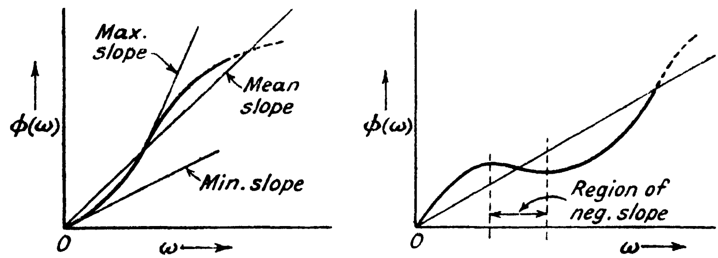
$$\left[\frac{\delta\phi(\omega)}{\delta\omega}\right]_{MAX} - \left[\frac{\delta\phi(\omega)}{\delta\omega}\right]_{MIN} \quad . \quad . \quad . \quad . \quad (210)$$

and the bulk of the response signal will be delayed by the time given by the mean slope $\delta\phi(\omega)/\delta\omega$. Thus serious distortion of the build-up time and of the widths of pulse signals can arise owing to bad departures of the phase-shift characteristic from the ideal straight line (see Fig. 63 (a)).

The general effects of phase distortion on a step wave or step-modulated carrier wave signal are a virtual delay and a spoiling of the build-up time, sometimes accompanied by overswings. The examples of Fig. 51 show these effects, though in this example amplitude distortion, here comprising a gradual attenuation of high-frequency components, is playing some part. Since the build-up of the response must take some finite time no exact figure may be given to the delay time of the whole wave, though a safe approximation may be made by taking the time to the point at half steady-

state amplitude.* This is often sufficient for practical purposes, provided that (1) the phase-shift curve does not vary from the linear very rapidly with frequency, i.e. when $d^2\phi(\omega)/d\omega^2$ is small, and (2) the amplitude characteristic is fairly uniform.

The exact shape of the build-up curve and the overswings, if any, depend on the exact shape of the phase-shift characteristic. If this characteristic is concave (positive curvature) then the higher frequencies are more delayed than the lower, which has the effect of spoiling the build-up time but reducing the overswings (sometimes given the descriptive name, "drooling" build-up), whereas a convex characteristic means the higher frequencies are less delayed than the lower and the response build-up can be steeper, with appreciable overswings. Some networks may have a phase-shift




(a) Variations in $\delta\phi(\omega)/\delta\omega$. (b) Region or Negative $\delta\phi(\omega)/\delta\omega$.

Fig. 63.—Variations from the Linear Ideal of a Phase Characteristic.

characteristic with a *negative* slope, at least over part of the frequency range (Fig. 63 (b)). This does not imply that the response wave arrives before the applied wave, but that certain frequencies are relatively advanced,[7] since such a negatively sloping part must be superimposed on an average positively sloping characteristic, while at the same time the amplitude characteristic must have some influence.

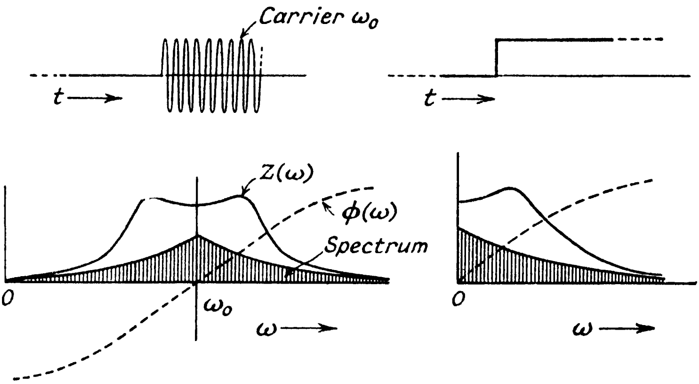
## 41. Response to an amplitude-modulated signal and response to its envelope

Where amplitude-modulated systems are concerned it is usually the case that the distortion of the envelope is of sole interest, since this contains the information in the signal, and in this connection there is a very important theorem which may be used to save

* This checks very well with the curves in Figs. 51 and 52.

considerable labour when calculations or approximations are made of the frequency distortion of a channel.

It has been shown in Chapter 2, Sec. 16, that the spectrum of an amplitude-modulated carrier is symmetrical about its carrier frequency, $\omega_0$, and that it is of the same form whatever the frequency $\omega_0$, *including zero*, so that the spectrum of the modulated wave is identical with the spectrum of its envelope, but shifted along the frequency axis. The theorem has been proved very simply for the case of sinusoidal modulation [20] and it must therefore be true for periodic and transient modulation, since it may be applied separately to every sinusoidal component in the envelope. (The theorem is illustrated by Figs. 18 and 19.) The same principle was applied in




(a) Step-modulated Wave Applied to a Band-pass Filter.

(b) Envelope Wave Applied to a Low-pass Filter.

Fig. 64.—Relation Between the Response to a Modulated Signal and Response to its Envelope.

Chapter 2 (see Fig. 20) to the vector diagram, resulting in the idea of a stationary carrier vector with pairs of sideband vectors rotating in opposite directions about its extremity.

Consider the application of an amplitude-modulated wave to a band-pass filter with ideally symmetrical frequency characteristics about a mid-band $\omega_0$. For example, Fig. 64 (a) shows a step-modulated wave and, below, its spectrum superimposed on a set of band-pass filter characteristics. The response wave will also have a perfectly symmetrical spectrum, and the carrier $\omega_0$ may therefore be "reduced to zero" frequency. The envelope of this response wave could therefore have been obtained by applying the spectrum of the modulated wave's envelope (i.e. the step wave of

Fig. 64 (b)), to an equivalent set of low-pass (i.e. envelope-frequency) characteristics. These equivalent characteristics are simply the band-pass characteristics transferred down to zero mid-band frequency.

This theorem is only approximately true, since some asymmetry is inevitable in all band-pass characteristics, even those of constant-$K$ type (see Sec. 30), though its accuracy will be found to be sufficient for most practical cases. If the carrier $\omega_0$ does not coincide with the mid-band frequency of the filter the theorem cannot be applied directly, since the response-wave spectrum cannot be symmetrical. Such cases may, however, be dealt with by a modification of the theorem, which is the subject of Chapter 7.

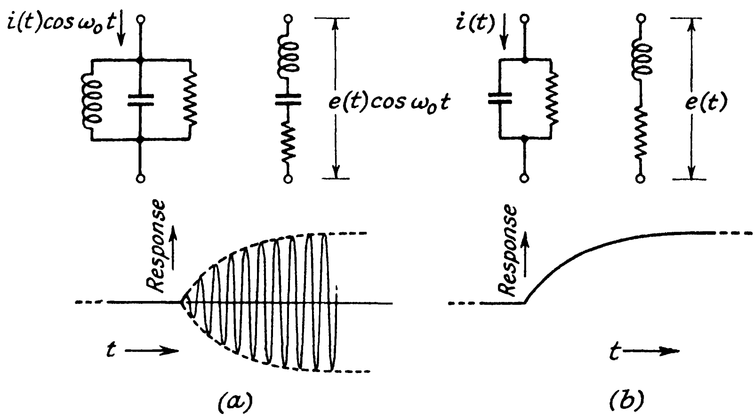



Fig. 65.—Response of Equivalent (a) "Band-pass" and (b) "Low-pass" Circuits to Step-modulated Carrier Wave and Step Wave Respectively.

The reader should note that this equivalence of characteristics has nothing to do with the fact that *physical structures* may be produced having such equivalent characteristics (the low-pass and band-pass equivalents of Sec. 33), though it does show that band-pass and low-pass filters may be constructed that give identical distortion, within reasonable limits of accuracy, to the envelope of a modulated carrier and to the envelope wave itself, respectively.

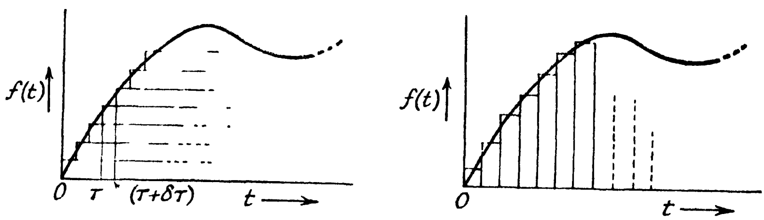
From this theorem it follows that any calculations which may have been made on the response of a low-pass filter may give directly the envelope distortion of the equivalent band-pass filter.

For example, Fig. 65 shows the response of (b) a parallel $RC$ circuit to a step wave of current and (a) the response of its

"band-pass equivalent" tuned to $\omega_0$ when excited by a step-modulated carrier, $\omega_0$. This result will be very nearly exact if the Q of the tuned circuit is fairly high (i.e. its band-width relatively narrow). These curves will also be the responses of the corresponding dual networks (see Sec. 28) shown in the same diagram, when excited by identical *voltage* waves.

## 42. The Superposition Theorem—complex transient response by the addition of simple responses

The response of a linear network to a transient having a complex waveform, perhaps not expressible analytically, may be derived by considering the transient to consist of a number of step waves of suitable amplitudes and times of starting (Fig. 66 (a)). The step wave is a basic waveform in circuit transient analysis and from this




(a) Division into Step Waves. (b) Division into Pulse Waves.

Fig. 66.— Use of Superposition Theorem.

point of view it is the only one which need be used, all others being derived by suitable addition. However we shall see later that there can be advantages in using a narrow pulse type of waveform, dividing up a transient as shown in Fig. 66 (b), and that there is a simple relation between these two types.

An approximation is involved in either case, if steps of finite amplitude or pulses of finite width be used, though by making these finite quantities sufficiently small any desired accuracy may be obtained in practical calculations. The principle may be applied directly to the envelopes of amplitude-modulated waves by use of the theorem outlined in the last section, and responses of symmetrical band-pass networks may be calculated by determining the response of the equivalent low-pass characteristics to a step or pulse wave, and then transferring this envelope response to the mid-band carrier frequency.

We may express this addition of step-wave responses as a

mathematical operation, thus extending the process from a numerical approximation to an exact method of analysis. Suppose that it is required to calculate the response of a circuit to a transient $f(t)$ of known analytical form and that the response $h(t)$ to a step wave of known amplitude (unit amplitude) has already been determined. With reference to Fig. 66, consider the response to that particular step wave commencing at a time $t=\tau$. The amplitude of the transient $f(t)$ at this point is $f(\tau)$, and therefore:

$$\text{Amplitude of step}=\frac{df(\tau)}{d\tau}\,.\,\delta\tau \text{ units}$$

starting at time $t=\tau$, where $\delta\tau$ is the time interval between successive step waves. The response due to a step wave of this amplitude is:

$$\text{Single step response}=f'(\tau)\,.\,h(t-\tau)\,.\,\delta\tau \quad . \quad (211)$$

(where $f'(\tau)=df(\tau)/d\tau$) at any time $t$ after this particular step wave has been applied. Hence the resultant response, at any time $t$, due to the combined effect of all the step waves, is:

$$\text{Resultant response}=\sum_{\tau-0}^{\tau=t} f'(\tau)\,.\,h(t-\tau)\,.\,\delta\tau \quad . \quad . \quad (212)$$

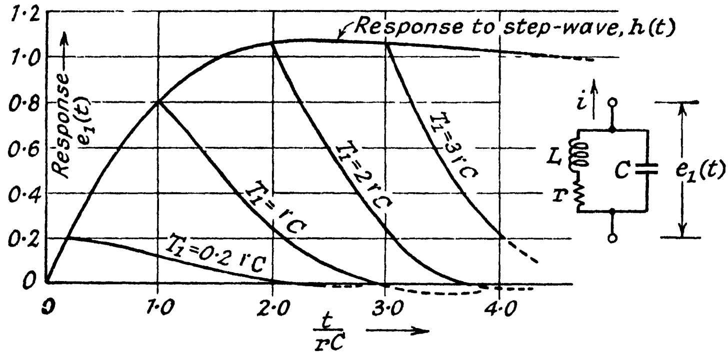
If the interval $\delta\tau$ is made indefinitely short, the $\Sigma$ becomes an integral:

$$\text{Resultant response}=\int_0^t f'(\tau)\,.\,h(t-\tau)d\tau \quad . \quad . \quad (213)$$

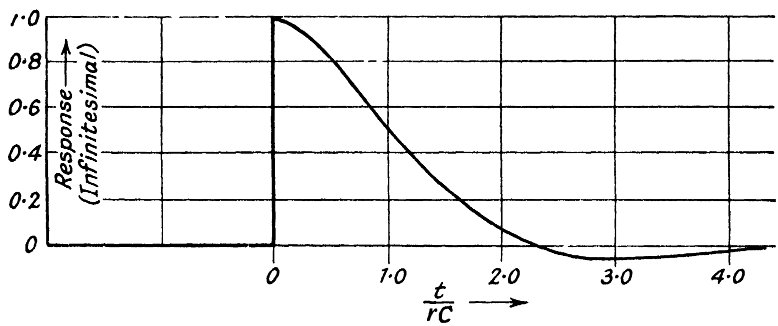
which is known as Du Hamel's integral, and which may be written in a number of other ways. If the response $h(t)$ of a circuit to a step wave has been calculated as an analytical expression then the result may be applied [9, 24] to find the response to other transients $f(t)$ of known analytical form, by the use of the above integral, provided that the integral can be evaluated. The variable $\tau$ is only a parameter of integration and vanishes when the limits of the integral are inserted.

The simplest application of this principle of Superposition is the evaluation of the response of a circuit to a rectangular pulse wave, of duration $T_1$, knowing its response $h(t)$ to a step wave, by considering the pulse wave to consist of two step waves of opposite sign displaced by a time $T_1$. Fig. 67 (a) shows an example, the voltage response $e_1(t)$ of a damped tuned circuit to a rectangular pulse of current, using several values of $T_1$ in terms of the circuit time-constant $rC$. These curves have been obtained by plotting the

step-wave response $h(t)$ (easily calculated in this instance by direct solution of a differential equation), then plotting the same response, inverted, a time $T_1$ later and adding the two curves. As a point of interest the circuit used here is a common form of anode load for television amplifiers, such as was given in Fig. 51 as an example (the circuit $Q=1/\sqrt{2}$). It can be seen from these curves that, as




(a) Response of Circuit to Pulse-wave Current of Duration $T_1$ Computed from the Step-wave Response. $[L/Cr^2=Q^2=0{\cdot}5.]$




(b) Response of Circuit to an Indefinitely Short Pulse, $T_1 \to 0$.

Fig. 67.—Relation Between Step-wave and Pulse-wave Response. Circuit $Q=1/\sqrt{2}$.

the applied pulse duration $T_1$ is reduced, the response becomes more and more non-symmetrical, having a finite time of rise and an infinite time of decay. This is a general effect in the response of circuits to narrow pulses, when the pulse duration becomes very short compared with the significant time-constants of the circuit.

In the limiting case of $T_1 \to 0$ the build-up time of the response will be instantaneous, though the response amplitude will be

infinitesimal. Fig. 67 (b) shows the response in this case, with the amplitude scale in arbitrary units.

This corresponds to a true impulse or "ballistic" excitation of the circuit, energy being supplied in an infinitesimal time, so that the circuit cannot change its condition before the excitation has ceased. The resulting response is therefore a "force-free," or natural, oscillation.

It may be shown that, apart from *magnitude*, the response waveform to the extremely narrow pulse is the differential coefficient of the step-wave response $h(t)$. Thus if the response to a step wave, applied at time $t=0$, is $h(t)$, the response to a negative step wave applied at time $t=T_1$ is $-h(t-T_1)$. Then the response to a rectangular pulse of duration $T_1$ is:

$$[\text{response to pulse}]=h(t)-h(t-T_1) \quad . \quad . \quad (214)$$

Now let $T_1$ become very short and write it as $\delta t$. Dividing both sides of (214) by $\delta t$:

$$\frac{1}{\delta t}\,.\,[\text{response to pulse}]=\frac{h(t)-h(t-\delta t)}{\delta t} \quad . \quad (215)$$

which in the limit $\delta t \to 0$ becomes the differential coefficient $dh(t)/dt$. The $\delta t$ on the left of 215 shows the infinitesimal magnitude of the pulse response.

Such responses may be met with in practice when extremely short pulses are used, since the exact differential form is very rapidly approached as $T_1$ becomes less than the circuit time constant. (Note the example in Fig. 67.)

## 43. Response to an extremely short pulse

An alternative to the transient method of regarding pulse response is the steady-state approach, and there are some interesting facts concerning the frequency spectrum of a circuit's response to a very short pulse.

The spectrum of a pulse of duration $\delta t$, approaching zero, is continuous and uniform from zero to infinite frequency; every term is cosine and has the same infinitesimal amplitude. In practice, if the pulse is very short compared with the circuit time-constants, its spectrum will be nearly constant over the whole range of frequencies passed by the circuit. This uniformity of spectrum may be seen, with reference to Fig. 29 (Chapter 2), by considering the conjugate spectrum of the rectangular pulse of duration $T_1$.

This spectrum falls to zero amplitude at frequencies $\pm 1/T_1$. As $T_1$ is reduced and finally made infinitesimal, these frequencies pass out to $\pm\infty$, so that the spectrum is of uniform amplitude over any finite range of frequencies. The non-uniformity of the spectrum over any band-width in a practical example with a finite pulse duration $T_1$ may be judged from these curves in Fig. 29, by comparing the band-width to the frequency $1/T_1$.

This continuous spectrum of cosine terms, having infinitesimal amplitudes, may be summed so that the pulse $I$ may be written:

$$\text{Indefinitely short pulse, } I=\int_{-\infty}^{+\infty}\cos\omega t\,.\,d\omega \quad . \quad . \quad (216)$$

and such a pulse has an indefinitely small energy.* If such a pulse of current be applied to the terminals of a network the response voltage waveform must have a spectrum identical with the frequency characteristics of the network (if a set of curves be plotted against a frequency scale there is no telling whether they represent the characteristics of some network or the continuous spectrum of a transient—they merely represent a distribution of sinusoidal terms of certain amplitudes and phases).

If the network characteristics be $|Z(\omega)|$ and $\phi(\omega)$, the response to the current pulse (216), applied at $t=0$, will be $v(t)$ where:

$$v(t)=\int_{-\infty}^{+\infty}|Z(\omega)|\,.\cos(\omega t-\phi(\omega))\,.\,d\omega \quad . \quad . \quad . \quad (217)$$

$$=\int_{-\infty}^{+\infty}[|Z(\omega)|\cos\phi(\omega)]\cos\omega t+[|Z(\omega)|\sin\phi(\omega)]\sin\omega t\,.\,d\omega \quad . \quad (218)$$

$$=v_c(t)+v_s(t) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (219)$$

as for equation 186, except that the applied spectrum $a(\omega)$ is here constant. Now the first term, $v_c(t)$, represents the response of the real part, $|Z(\omega)|\,.\cos\phi(\omega)$, of the network characteristics; also it is seen to be the Fourier integral of this real part and it is symmetrical about $t=0$ since it consists of cosine terms only. Similarly, $v_s(t)$ is the response of the imaginary part, $|Z(\omega)|\,.\sin\phi(\omega)$, of the network characteristics; it is also the Fourier integral of this part and is skew-symmetrical about $t=0$. This is a particular case of the general conclusions reached in Sec. 35, and illustrated by Fig. 53, regarding the symmetry relations between the responses of the real

* As written, this integral has no mathematical value; in conjunction with the discussion here, it should merely be taken to mean the sum of an infinite spectrum of terms of the form $\cos\omega t$.

and imaginary parts of a network's characteristics. Since there can be no response before the time $t=0$, these responses, $v_c(t)$ and $v_s(t)$, must be identical in shape but must cancel out for $t<0$. Thus we may conclude that in a physical network the impedance characteristics (direct or transfer) must be such that the Fourier integrals of their real and imaginary parts are identical in form but of even and odd symmetry respectively.

In Sec. 36 the example was chosen of the response of a parallel $RC$ circuit to a step wave of current, such as would occur in the tetrode amplifier stage of Fig. 54 with a step wave of voltage on the grid. Consider instead the response to an idealised pulse, which will be given by the Fourier integral of the impedance characteristics of this circuit. These characteristics are given by equation 197 and the pulse response is therefore:

$$\begin{aligned} v(t) &= \int_{-\infty}^{+\infty} \frac{R}{1+(\omega CR)^2} \cos \omega t d\omega + \int_{-\infty}^{+\infty} \frac{\omega CR}{1+(\omega CR)^2} \sin \omega t d\omega \\ &= v_c(t) + v_s(t) \end{aligned} \quad . \quad (220)$$

in terms of the separate responses of the real and imaginary parts. From a list of Fourier integrals (such as G. A. Campbell's,[8] pairs 444 and 445) these integrals will be seen to be identical for positive values of $t$, but equal and of opposite sign for negative values, $v_c(t)$ being symmetrical about $t=0$ and $v_s(t)$ skew-symmetrical. The actual values from Campbell's list are:

$$v_c(t) = v_s(t) = \tfrac{1}{2}\epsilon^{-t/RC} \text{ for } t \text{ positive,}$$

$$v_c(t) = -v_s(t) = \tfrac{1}{2}\epsilon^{t/RC} \text{ for } t \text{ negative,}$$

so that the responses prior to $t=0$ add up to zero.

Use of the above theorem may save labour, since in such transient calculations the Fourier integral of either the real or the imaginary part needs to be calculated, but not both.

If the response of a network to a pulse has been calculated in this way, the step-wave response is given by integration, graphical or otherwise, and the response to any other wave may be found by use of the Superposition Theorem.

We have referred here to a response *voltage*, but the relations that are proved here between the real and imaginary parts of the characteristics apply to any of the four forms of characteristic, with current or voltage drive and current or voltage response.

## 44. Successive differentials and integrals of a response wave

We shall end this chapter with another simple theorem which may be used to extend one completed transient calculation to a number of other cases.

The response of a network to the differential coefficient of a driving wave $f(t)$ is given by the differential coefficient of its response to the wave $f(t)$. This is true as regards the waveform *shape* of the differentials but not necessarily as regards their magnitudes.

We have already dealt with one particular case. In Sec. 42 it was demonstrated that the response of a network to an infinitesimal pulse is the differential of its response to a step wave, and this was illustrated by an example, Fig. 67 (a) and (b). Also the pulse signal is, apart from absolute magnitude, the differential coefficient of the step wave. The spectrum $b(\omega)$ of sine terms constituting a step wave is given by equation 193, so that any one term of any frequency $\omega$ is:

$$b(\omega) \sin \omega t = \left[\frac{I}{\pi}\right] \frac{\sin \omega t}{\omega} \quad . \quad . \quad . \quad . \quad (221)$$

Differentiating:

$$\frac{d}{dt}[b(\omega) \sin \omega t] = \left[\frac{I}{\pi}\right] \cos \omega t \quad . \quad . \quad . \quad . \quad (222)$$

which has an amplitude independent of frequency. Thus the wave corresponding to the differential coefficient of a step wave has a uniform constant amplitude spectrum of cosine terms, which is the spectrum of an infinitesimally short pulse (see Sec. 43). The theorem may be proved in the general case, for a linear circuit, by considering any one term $e(\omega)$ in the spectrum of the applied driving wave:

$$e(\omega) = A \sin (\omega t + \theta) \quad . \quad . \quad . \quad . \quad . \quad (223)$$

and

$$\frac{d}{dt}e(\omega) = A\omega \,.\, \cos (\omega t + \theta) \quad . \quad . \quad . \quad . \quad (224)$$

The responses of a network of impedance $|Z(\omega)|$, $\phi(\omega)$, to these two waves (223 and 224) are respectively :

$$|Z(\omega)| \,.\, A \sin (\omega t + \theta - \phi(\omega)) \quad . \quad . \quad . \quad (225)$$

and

$$|Z(\omega)| \,.\, A\omega \,.\, \cos (\omega t + \theta - \phi(\omega)) \quad . \quad . \quad . \quad (226)$$

and clearly equation 226 is the differential coefficient of 225, with respect to time. Since this is true for one frequency component it is true for any complex waveform, by the Superposition principle.

The theorem is conversely true as regards integration, and of course for any number of successive integrations or differentiations. The commonest application is the estimation of the step-wave response from the calculated pulse response by integration, graphical or otherwise. Frequently the question of the magnitude of the

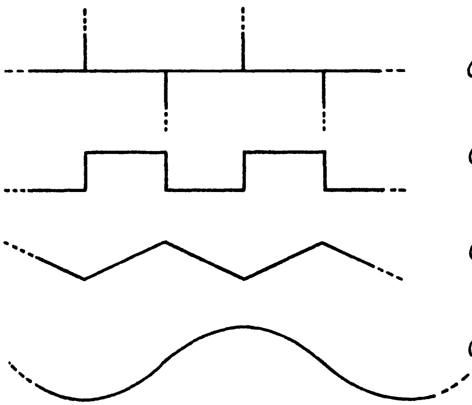



Fig. 68.—Successive Integration of a Waveform.

response may be settled by a consideration of the steady-state or the D.C. response.

Starting from the pulse response, successive integration gives the response to a step wave, a wave of linear build-up, a parabolic build-up, cubic and higher powers, as illustrated by Fig. 68.

## REFERENCES

1. A. C. Bartlett, "The Calculation of Modulation Products," Part I, *Phil. Mag.*, Vol. xvi, Oct. 1933; Part II, Vol. xvii, March 1934.
2. D. C. Espley "Harmonic Production and Cross-modulation in Thermionic Valves with Resistive Loads," *Proc. I.R.E.*, Vol. 22, No. 6, June 1934.
3. K. Kupfmuller, "Uber Einschwingvorgange in Wellenfiltern," *Elec. Nacht. Tech.*, Vol. 1, Nov. 1924.
4. R. V. L. Hartley, "Transmission of Information," *Bell Syst. Tech. Journ.*, Vol. 7, p. 535, Aug. 1928.
5. W. L. Everitt, "Communications Engineering," McGraw-Hill, New York, 1937, Chap. 1.
6. J. Jeans, "Science and Music," C.U.P., 1937.
7. A. T. Starr, "Electric Circuits and Wave Filters," Isaac Pitman & Sons, 1938, 2nd ed., Chap. 12.
8. G. A. Campbell, "The Practical Application of the Fourier Integral," *Bell Syst. Tech. Journ.*, Vol. 7, p. 639, Oct. 1928.

9. G. W. Carter, "The Simple Calculation of Electric Transients," C.U.P., 1944.
10. A. V. Bedford and G. L. Fredendall, "Transient Response of Multi-stage Video-frequency Amplifiers," *Proc. I.R.E.*, Vol. 27, No. 4, April 1939, p. 277.
11. A. V. Bedford and G. L. Fredendall, "Analysis, Synthesis, and Evaluation of Transient Response of Television Apparatus," *Proc. I.R.E.*, Vol. 20, Oct. 1942.
12. S. P. Meade, "Phase Distortion and Phase Distortion Correction," *Bell Syst. Tech. Journ.*, Vol. 7, p. 195, April 1928.
13. A. A. Barco, "Measurement of Phase Shift in Television Amplifiers," *R.C.A. Review*, Vol. III, No. 4, April 1939.
14. G. W. Edwards and E. C. Cherry, "Amplifier Characteristics at Low Frequencies, with Particular Reference to a New Method of Frequency Compensation of Single Stages," *J.I.E.E.*, Vol. 87, No. 524, Aug. 1940.
15. H. E. Kallman, "Portable Equipment for Observing Transient Response of Television Apparatus," *Proc. I.R.E.*, Vol. 28, Aug. 1940.
16. R. V. L. Hartley, "Steady-state Delay Related to Aperiodic Signals," *Bell Syst. Tech. Journ.*, Vol. 20, p. 222, April 1941.
17. H. E. Kallman, R. E. Spencer, and C. P. Singer, "Transient Response," *Proc. I.R.E.*, Vol. 33, No. 3, March 1945.
18. J. R. Carson, "The Building-up of Sinusoidal Currents in Long, Periodically Loaded Lines," *Bell Syst. Tech. Journ.*, Vol. 3, p. 558, Oct. 1924.
19. H. Nyquist and S. Brand, "Measurement of Phase Distortion," *Bell Syst. Tech. Journ.*, Vol. 9, p. 522, July 1930.
20. R. V. L. Hartley, "Relations of Carrier and Side-bands in Radio Transmission," *Bell Syst. Tech. Journ.*, Vol. 2, p. 90, April 1923.
21. W. L. Sullivan, "Analysis of Systems with Known Transmission-Frequency Characteristics by Fourier Integrals," *Electrical Engineering*, Vol. 61, No. 5, May 1942.
22. R. G. Schiffenbauer, "Phase Distortion in Television," *Wireless Engineer*, Vol. 13, p. 21, Jan. 1936.
23. D. G. Tucker, "Transient Response of Filters," *Wireless Engineer*, Vol. 23, No. 269, Feb. 1946, p. 36, and March 1946, p. 84.
24. V. Bush, "Operational Circuit Analysis," John Wiley, New York (Chapman & Hall, London), 1936.

The following are also of interest:

(A) H. S. Carslaw, "Introduction to the Theory of Fourier Series and Integrals," Macmillan & Co., London, 1930.
(B) C. H. Smith, "Amplification of Transients," *Wireless Engineer*, Vol. 10, p. 296, June 1933.
(C) D. M. Johnstone, "Inductively Compensated Tetrode Amplifiers," *Wireless Engineer*, Vol. 15, p. 208, April 1938.
(D) O. S. Puckle, "The Transient Aspect of Wideband Amplifiers," *Wireless Engineer*, Vol. 12, p. 251, May 1935.
(E) H. W. Bode, "Relations between Attenuation and Phase in Feedback Amplifier Design," *Bell Syst. Tech. Journ.*, Vol. 19, No. 3, July 1940.
(F) H. J. Josephs, "Heaviside's Electric Circuit Theory," Methuen Monograph, London, 1946.

# CHAPTER 5

# Some Approximations and Idealised Results

## 45. The need for approximation

It is only too frequently the case that the use of the Fourier integral method of calculating the response of a circuit to a transient signal may lead to extremely difficult or insoluble integrals, although the method itself, as an idea, is very simple. The limitation is purely a mathematical one, and engineering principles are not involved. This Fourier integral approach to the problem is, however, of extreme importance, not only as a way of looking at the general problem of transient response but also for practical application by the use of what may be termed *idealised* conditions; such conditions involve the assumption of circuit properties which are non-physical but which may be justified by the fact that they lead to comparatively simple calculations and give sufficiently accurate results for practical use, provided that their non-physical nature is borne in mind and any limitation due to this fact taken into account.

It is often unnecessary, in communications problems, to make exact transient response calculations, but, instead, general points concerning band-width or build-up time need to be settled or the question arises of interference signals set up in one channel due to transmission through a neighbouring channel. In such cases the driving signal may be considered to consist of a small step or short pulse, comprising the smallest piece of detail of the signal ultimately to be transmitted by the system. The use of such steps or pulses does, as we have seen in the last chapter, simplify the calculations, but does not assume anything non-physical about the system—except, it may be argued, that perfect steps and infinitesimal pulses cannot be produced by physical circuits. The point is important in that it illustrates the difference between an approximation and an idealised result—the step or pulse can be reproduced by physical circuits *as accurately as we please* by taking sufficient trouble. Similarly with other approximations we have discussed already; the use of band-pass/low-pass equivalent circuits, the use of a Fourier series instead of an integral, the use of the Superposition

Theorem for complex transient computations, or the transfer of a carrier frequency to zero, all these theorems can be made true to any desired accuracy. Other types of approximation are commonly used in communications. For example resistance is usually neglected in the design of filters and frequently they are assumed to be iteratively terminated, so that the practical circuit has characteristics which differ from those of the basic structures as listed in text-books, but nevertheless may be made to approach the basic characteristics as closely as desired. Theoretically speaking there are no physical limitations, and all these simplifications involve merely a question of degree.

However, if response calculations are simplified by the assumption of network frequency characteristics which cannot be achieved physically, the result will be truly idealised. Thus certain geometric shapes of characteristics can be used in order to obtain soluble Fourier integrals, and very rapid estimates may be made of the relations between band-width and build-up time and of other factors, but, as will be seen from the more detailed examination in the next section, the response signals of such idealised systems may show a positive value before the time $t=0$ (when the input signal is first applied): clearly a physically impossible result. The conclusion may be drawn that such geometric shapes of characteristics cannot be obtained by any structure of electric elements. The shapes of the characteristics of physical structures are bound by certain laws, and we cannot just choose any shape for its convenience and expect it automatically to obey these laws. Thus, provided that no rash assumptions are drawn from the fact that idealised characteristics give a finite response before $t=0$, these characteristics are extremely useful in practice for approximate calculations.

## 46. Band-width and build-up time—the idealised filter response

The band-width of any practical filter is an indefinite quantity, since all such networks show some kind of a response up to "infinite" frequency. From the point of view of transient response calculation there is little reason why the theoretical filter cut-off frequency should be chosen to be the exact band-width boundary, although the attenuation of a filter must increase rapidly in the neighbourhood of this frequency, even if it consists of one or of a few sections only, non-iteratively terminated. For example, the theoretical cut-off frequency of the constant-$K$ filter section of Fig. 33 is at $\omega CR=1{\cdot}0$.

There are also many structures in common use which are not proper filters and which have no "cut-off" frequency.

If a nominal cut-off is to be chosen, for idealised calculations, there is good reason for taking it at the frequency $\omega_1$, such that the area lying underneath a uniform amplitude characteristic up to this frequency, equals the total area under the practical characteristic from zero to infinite frequency. That is, with reference to Fig. 69:

$$\omega_1 = \frac{1}{Z(0)} \int_0^\infty |Z(\omega)| d\omega \quad . \quad . \quad . \quad . \quad . \quad (227)$$

These characteristics have been drawn in this figure in their conjugate form so as to be applicable to low-pass or band-pass circuits. The practical characteristics are shown dotted and the idealised ones by

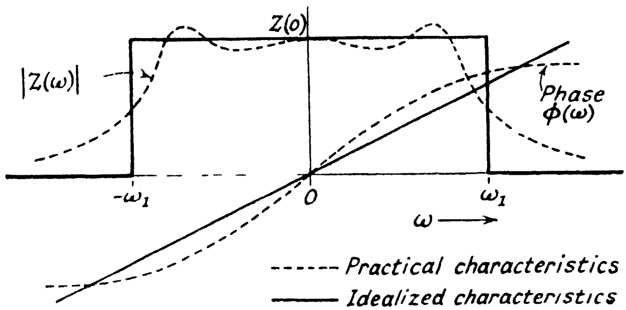



Fig. 69.—Equivalent Idealised (Flat-topped) Characteristics.

solid lines, the idealised amplitude characteristic being uniform up to the frequency $\omega_1$ and phase shift being of uniform slope, $t_1 = \phi/\omega$ approximately equal to the average slope of the practical phase-shift characteristic. If the phase-shift characteristics are neglected these two amplitude characteristics, the practical and the idealised, are such that their responses to an applied step wave will have the same maximum rate of build-up. This may be proved as follows: the sine-wave components, $b(\omega) \sin \omega t$, of a step wave are all in phase at $t=0$, corresponding to the mid-point of the step, and all have their steepest slope at that point. Now the slope (i.e. rate of rise) of a sine wave at $t=0$ is equal to the product of the frequency and amplitude:

$$\frac{u}{dt}[b(\omega) \sin \omega t] = [\omega \, . \, b(\omega)] \cos \omega t \qquad (228)$$

which, at $t=0$, is equal to $\omega \, . \, b(\omega)$. The maximum rate of rise of

the sine-wave components of the response signal of a network with a characteristic modulus $|Z(\omega)|$, when this step wave is applied to its input terminals, will then be $|Z(\omega)| \,.\, \omega \,.\, b(\omega)$. However, the component amplitudes $b(\omega)$ in a step wave of amplitude $I$ are equal to $Id\omega/\pi\omega$ (equation 196), so that this maximum rate of rise is proportional to $|Z(\omega)|$.

Summing these rates, since each component contributes equally to the maximum rate of rise of the resultant response, this rate is given by:

$$\text{Maximum slope of response to step wave} = \frac{I}{\pi}\int_0^\infty |Z(\omega)|d\omega \quad . \quad (229)$$

which is $I/\pi \times$ the area under the network modulus characteristic.

Furthermore, in the case of the flat-topped idealised filter, the area (and hence the maximum slope of the response) is proportional to $\omega_1$ the band-width (this is approximately true also in practice for reasonably flat filters). Thus, their areas being equal, the responses of the idealised and the practical characteristics in Fig. 69 will have the same maximum rate of rise, though the response waveforms will differ in detail. The shape of the response of the idealised flat-topped characteristic [1] may easily be calculated and serves as a guide to the response of any physical network characteristic which it idealises.

The spectrum of a step wave of amplitude $I$ is given by equation 194:

$$f_s(t) = \frac{I}{2} + \frac{I}{\pi}\int_0^\infty \frac{\sin \omega t}{\omega} d\omega \quad . \quad . \quad . \quad [(194)]$$

and the idealised filter will pass all these spectrum components without distortion, but delayed by the time $t_1$, up to the frequency $\omega_1$, but will reject all others above this limit. The response wave $h(t)$ will thus be:

$$h(t) = \frac{E}{2} + \frac{E}{\pi}\int_0^{\omega_1} \frac{\sin \omega \overline{t - t_1}}{\omega} d\omega \quad . \quad . \quad . \quad (230)$$

where $E/2$ is the D.C. component. $E = I \,.\, Z(0)$ sets the absolute level of the response.

The integral above is an extremely important one, and although it cannot be evaluated as an algebraic expression it is very simple to use for the type of calculation we are considering. It is called

the "sine integral," frequently written $Si(\omega_1)$, and is tabulated[2, 3] against values of its upper limit $\omega_1$.

This response $h(t)$ (230) is plotted in Fig. 70 against a time scale $\omega_1 t$ so that it is "universal" and gives the response of such an idealised low-pass filter with any cut-off frequency $\omega_1$. Several points about this response curve are noteworthy. First, the oscillations extend to infinity in both positive and negative time directions, so that, no matter what time delay $t_1$ is assumed, such an idealised filter shows a definite signal response *before* the step wave is applied to its input terminals, clearly a physical impossibility. In Fig. 70 the time delay $t_1$ is reduced to zero since there is no point in choosing any arbitrary value. The physical impossibility of the

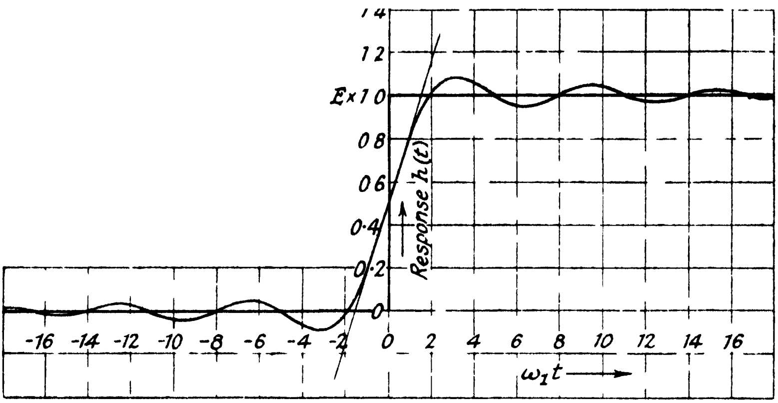



Fig. 70 Response of the Idealised Flat-topped Filter of Fig. 69 to an Applied Step-wave.

response before $t=0$ is of no consequence since it merely illustrates the non-physical nature of the frequency characteristics, a point we have already discussed in the preceding section. Secondly, and perhaps more important, the build-up time of this low-pass filter of band-width $\omega_1$ is $\pi/\omega_1$ if measured between the points at which a tangent at the origin cuts the zero and steady-state levels, that is to say the time in which the wave would rise to its full steady-state value if the sloping edge were to maintain its maximum rate of rise, which occurs at the origin. This maximum rate of rise is, as we have already seen, the same as the maximum rate of rise of the response of a practical filter whose characteristic encloses the same area as the idealised one (Fig. 69); it is given by equation 229. In this idealised case, since the characteristic encloses an area $|Z(0)| \cdot \omega_1$

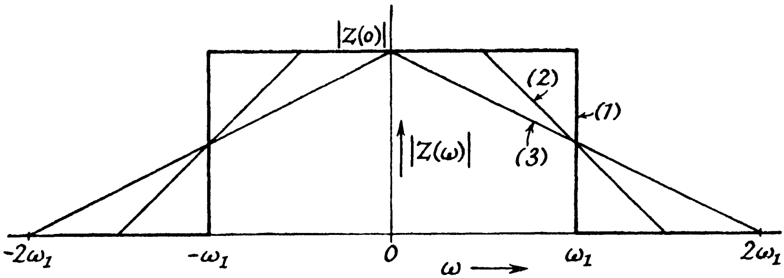
the maximum rate of rise of the response (Fig. 70) is $I \,.\, |Z(0)| \,.\, \omega_1/\pi$. But $I \,.\, |Z(0)|$ is the steady-state value $E$ of the response. Thus the maximum rate of rise is $E\omega_1/\pi$ so that the time of rise to this value $E$ is $\pi/\omega_1$, as we have already observed.

The reader should note, for the purpose of rapid build-up calculations, that this straight-line build-up time $\pi/\omega_1$ is equal to one half period of the cut-off frequency $\omega_1$, whereas the build-up time of the actual idealised response wave, measured between the lowest and highest points of the sloping edge, is $2\pi/\omega_1$, or one whole period of the cut-off frequency, and the curve of build-up is approximately sinusoidal in shape. The third point about this response wave is that it is skew-symmetrical about $t=0$ (assuming zero delay), being a sine function, and the overswing is 8·6 per cent. greater than $E$, the steady-state value.

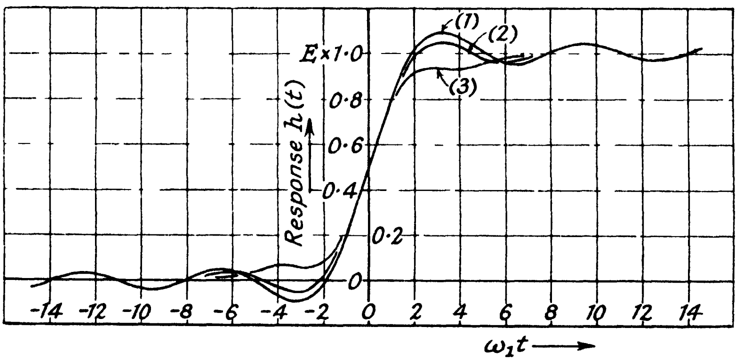
Finally, the frequency of these response overswings is seen to be approximately equal to the cut-off frequency $\omega_1$. This particular amount and frequency of overswing corresponds to this one type of idealised filter, and other filters having slower rates of cut-off have correspondingly less overswing on their transient responses. A fairly accurate idea of the amount and frequency of overswing in any particular case (phase distortion excluded) may be obtained by considering other types of idealised filter characteristics, for instance those shapes plotted in Fig. 29 (a). These curves in these figures were used in Chapter 2 for illustrating pulse waves and their spectra, and their reciprocity, as proved by Fourier transform theory; however, if the curves (a) be taken as frequency characteristics of idealised filters, the curves (b) will represent their pulse responses* (see Sec. 43, Chapter 4), from which the step-wave responses may be found by graphical integration. The trapezium and triangular shapes are most convenient for the present purpose and such idealised filter shapes are shown in Fig. 71 (a) together with (b) their step-wave responses. These responses, together with the rectangular filter response, illustrate the effect of a reduced rate of cut-off and by interpolation may give a closer approximation to the response of an actual filter in many practical cases. All these idealised responses are skew-symmetrical and show a finite amplitude before $t=0$. This symmetry is upset in practical filters by the phase-shift characteristic, which must be such as to remove all signs of a response before $t=0$, but which may increase the amount of overswing.

* Note we are using the idea of a Fourier *transform* here, having changed over the time and frequency scales in Fig. 29 (a) and (b).

These responses show clearly the large tolerance that is permissible in an amplitude characteristic before the transient response is seriously affected. Although these idealised characteristics have equal areas and equal band-widths at half-amplitude points, their shapes are very different. Nevertheless the three responses differ only in small details, mainly in the shape of the first overswing.




(a) Idealised Filter Characteristics.




(b) Responses of the Idealised Characteristics (1), (2) and (3).

Fig. 71.—Idealised Low-pass Filters and Their Step-wave Responses.

In general the response overswings are reduced as the rate of cut-off of the filter characteristic is lessened.

These idealised responses, Fig. 71 (b), may be taken as the envelopes of the responses of the equivalent *band-pass* filters formed by transferring the axis of symmetry of the idealised characteristics of Fig. 70 (a) up to any carrier frequency $\omega_0$. The carrier frequency of the responses will be $\omega_0$, in accordance with the theory outlined in Sec. 41, and the applied signal in this case must be a carrier $\omega_0$ modulated by a step wave. The build-up time of the envelope will

again be $\pi/\omega_1$ in the case of the rectangular filter characteristic, if the band-width is $\pm\omega_1$, as is illustrated by Fig. 72.

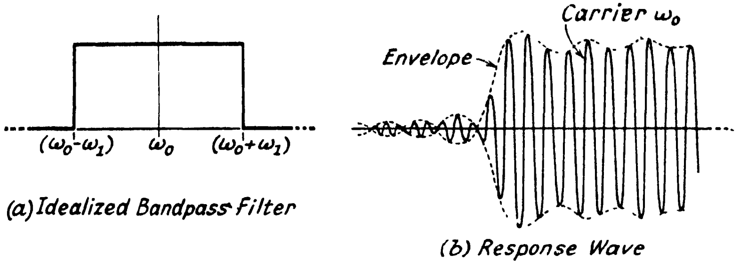



Fig. 72.—Response of Idealised Band-pass Filter to Step-modulated Carrier Wave $\omega_0$ (Accurate Envelope Form in Fig. 70).

## 47. Avoiding the response before $t=0$

An alternative way of calculating the response of the idealised filter of Fig. 69 has been suggested by D. A. Bell [4] which avoids the anomaly of a finite response before the step wave is applied to the input terminals at $t=0$ and which gives idealised responses more closely approaching the responses of practical filters. This alternative depends on the fact that a step wave may be represented by more than one mathematical expression, in the following manner.

The expression 230 for the idealised filter response was derived from the form of step wave given by equation 194 in Sec. 35. This form was developed by considering the repeated square wave of period $T_0$ (see Fig. 25, with the time origin at point (2)) and letting $T_0$ become infinite. The response, plotted in Fig. 70, was obtained by limiting the spectrum of the step wave to the frequency $\omega_1$ (the limit of the integral in equation 230) and this response already has the value $0{\cdot}5E$ at $t=0$, however high the frequency $\omega_1$, even if this limit becomes infinite. If $\omega_1$ approaches infinity this response will approach the shape of the applied step wave and hence this step wave must be assumed to have the value 0·5 at the point of discontinuity $t=0$.

However, an alternative expression for the step wave may be derived from the double step-wave form of duration $T_0$, Fig. 73 (which has zero average value or D.C. component). By allowing $T_0$ to become infinite, the expression for the step wave $f_s(t)$ becomes [3]:

$$f_s(t)=\frac{2}{\pi}\int_0^\infty \frac{\sin \omega t}{\omega}d\omega \quad . \quad . \quad . \quad . \quad . \quad (231)$$

A further result of this derivation is that the expression 231 is valid only for $t>0$, while $f_s(t)=0$ for $t<0$.

The response $h(t)$ of the idealised filter of Fig. 69 in this case is

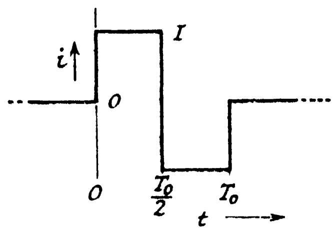



Fig. 73.—An Alternative Waveform from which a Step-wave may be Developed.

obtained by limiting the above spectrum to $\omega_1$ and inserting the constant phase delay $t_1$:

$$h(t)=\frac{2E}{\pi}\int_0^{\omega_1}\frac{\sin \omega\overline{t-t_1}}{\omega}d\omega \text{ for } t>0 \text{ only} \quad . \quad . \quad (232)$$

which is similar to 230 except that its value is zero at $t=0$ and for all values of $t<0$. This response is plotted in Fig. 74 and it should

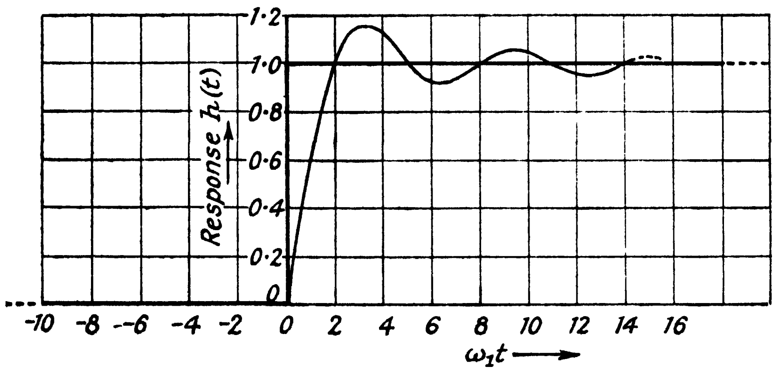



Fig. 74.—Response of the Idealised Flat-topped Filter, Using the Expression (232) for the Step-wave.

be compared with the response in Fig. 70 obtained by the other expression for a step wave (194). If the band-width $\omega_1$ is allowed to become infinite this response will still maintain its value zero for $t<0$.

This form of idealised response is therefore much nearer to physical reality and D. A. Bell has indicated that it is identical with

the response of a constant-$K$ filter non-reflectively terminated when a voltage step wave is applied (mid-series) to its input terminals. Such termination would require an infinite nümber of filter sections, but it is nevertheless a *physical* structure.

## 48. Response of idealised filters to short pulses

From the responses of these idealised filters to step waves, their responses to rectangular pulse-type waves of any duration may be computed by the Superposition Theorem, as explained in Sec. 42. The response to a rectangular pulse of duration $T_1$ is obtained by adding to the step wave response $h(t)$ the inverted response $-h(t)$ delayed by the time $T_1$. Furthermore, Sec. 42 showed that as the pulse duration $T_1$ became very short the waveform of the response tended to assume a fixed shape, until ultimately, as $T_1 \to 0$, the response to the infinitesimal pulse had a shape corresponding to the differential coefficient of the step-wave response.

The "shortness" of the pulse, $T_1$, was judged in Sec. 42 in relation to the time constants of the circuit being pulsed, but in our idealised filter characteristics there is no question of a time-constant. In this case our criterion of shortness is given by the spectrum of the pulse. If the pulse is so short that its spectrum is uniform over the idealised band-width the calculated response will be a true pulse response. There is no particular point in deducing the response to rectangular pulses of various finite durations for the idealised filter owing to there being some doubt concerning the exact form of the build-up curve and the exact amount of overswing, and such responses would largely depend on these factors for their waveshapes. However, the response to the infinitesimally short pulse is of theoretical interest in that it represents the ultimate signal definition of a channel of a given band-width. Since the spectrum of such a pulse is of uniform amplitude at all frequencies the energy contributed by every component to the resultant response will be equal. Equation 216 expresses the infinitesimal pulse, $I$, in terms of its uniform spectrum, and if this be applied to the idealised flat-topped filter of Fig. 71 (a), curve (1), the spectrum components will be reduced to zero outside the frequency range $\pm\omega_1$. Then:

Response of idealised filter to infinitesimal pulse =

$$|Z(0)|\int_{-\omega_1}^{+\omega_1}\cos \omega t d\omega \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (233)$$

which is
$$\propto \frac{\sin \omega_1 t}{t} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (234)$$

The actual magnitude of this response will be infinitesimal, though 234 gives its wave *form*. This waveform function is the Fourier integral of the flat-topped characteristic (see Sec. 43) and it is plotted in the charts of Fourier transforms (Fig. 29). From this same figure the pulse responses of the idealised trapezium, triangular and other forms of characteristic may be seen, since these pulse responses are always given by the Fourier transforms of the corresponding characteristics. It is interesting to see the slight changes in the pulse responses, Fig. 29 (b), with different characteristics (a), and to notice the effect of increasing the rate of cut-off, of having sharp corners, etc., on these idealised characteristics. Thus the response of the flat-topped (rectangular) characteristic has the largest overswing, while the "cosine-squared" shape and the triangular characteristics exhibit responses with almost negligible overswing.

It should be remembered that these pulse responses are given by the differential coefficients of the corresponding step-wave responses (see Sec. 44) so that these differences in the idealised characteristics may be interpreted in terms of their responses to step waves. By such a study of idealised characteristics we can see what factors control certain forms of signal distortion and to what degree, and we can then apply these observations to practical networks. Their use helps in analysing the contributions made by different parts of an energy spectrum to the eventual shape of a signal element, an operation which is of particular interest in the shaping of television signals (for correction of scanning aperture distortion, etc.).

## 49. The effect of a tailing frequency characteristic

All the idealised characteristics that we have considered so far have been limited in extent on the frequency axis, a condition which is impossible in practical networks. The general effect of the gradual tailing-off of a frequency characteristic is a reduction in the response overswings and a more gradual, smoother, build-up curve. The triangular characteristic in Fig. 71, curve (3), illustrates this point, although this characteristic does not extend beyond $2\omega_1$. The characteristics plotted in Fig. 75 (a) are two examples of idealised filters with skirts extending to infinity, and (b) shows their step-wave responses.

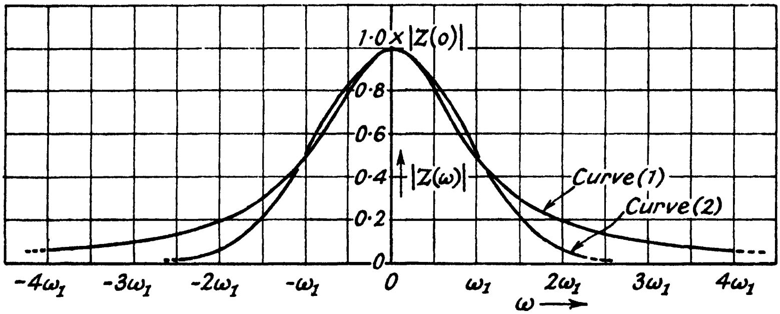
In this figure, curve (1) is given by the expression:

$$|Z(\omega)| = |Z(0)| \cdot \frac{1}{1+(\omega/\omega_1)^2} \quad . \quad . \quad . \quad . \quad (235)$$

and curve (2) by:

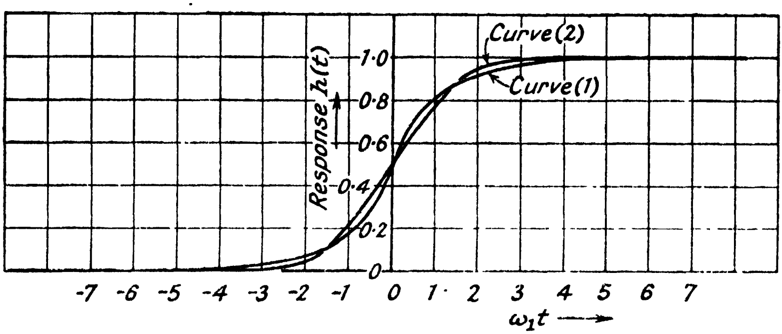
$$|Z(\omega)| = |Z(0)| \cdot \epsilon^{-(\omega/k\omega_1)^2} \quad . \quad . \quad . \quad . \quad (236)$$

where $k = 1{\cdot}195$, so that $\omega_1$ is the frequency at half-amplitude.




(a) Idealised Filter Characteristics with Tail. Curve (1), $|Z(\omega)| = \dfrac{|Z(0)|}{1+\left(\dfrac{\omega}{\omega_1}\right)^2}$.

Curve (2), $|Z(\omega)| = |Z(0)| \cdot \epsilon^{-(\omega/k\omega_1)^2}$—the "probability function."




(b) Response to Step-wave.

Fig. 75.—Response of Idealised Filters Having Tailing Characteristics.

The first of these curves has the shape of the real part of a damped tuned circuit characteristic (see Sec. 36, equation 197), and its Fourier integral, which gives the pulse response, is readily calculated as in Sec. 36. It is the exponential pulse:

$$\begin{cases} v(t) = \epsilon^{-\omega_1 t} \text{ for } t \text{ positive} \\ \quad\;\; = \epsilon^{\omega_1 t} \text{ for } t \text{ negative} \end{cases}$$

The step-wave response $h(t)$, as plotted in Fig. 75, is given by the integral of this pulse response.

The second of these idealised filter curves is of the "probability" [2, 3, 7] form and its Fourier integral [5, 6] is given by the same curve* (a self-reciprocal Fourier Transform), this curve being the infinitesimal pulse response:

$$v(t) \propto \epsilon^{-(\pi k \omega_1 t)^2} \text{ for } t \text{ positive and negative.}$$

The integral of this response again gives the step-wave response $h(t)$ of this idealised filter:

$$h(t) = \int_{-\infty}^{+\infty} \epsilon^{-(\pi k \omega_1 t)^2} dt \quad . \; . \; . \; . \; . \quad (237)$$

The steady-state value of $h(t)$, for large values of $t$, has been made unity in the figure. Values of this integral are to be found in tables, from which curve (2), Fig. 75 (b), has been plotted.

Both these responses tail off to infinity in both positive and negative time directions and have smooth build-up curves of slightly different forms, skew-symmetrical about $t=0$.

By the Superposition Theorem we may add together various idealised characteristics to give a composite idealised filter characteristic and the response of this is obtained by adding together the individual responses, bearing in mind the relation between the various time and frequency scales.

## 50. Shaping of a signal (picture) element

Summarising the results obtained so far, we may say that the responses of idealized filters with limited band-widths show overswings, the degree of overswing increasing with the rate of cut-off, whereas the responses of filters with drooping characteristics, tailing-off gradually towards infinite frequency, show a smooth build-up with no overswings. All kinds of intermediate response curves could be derived by adding together these various idealised characteristics in different ways, but this would profit us little. The idealised filters we have dealt with are sufficient to form a guide as to the approximate character of a response in a practical case—in which, we must remember, phase distortion is inseparable from amplitude distortion.

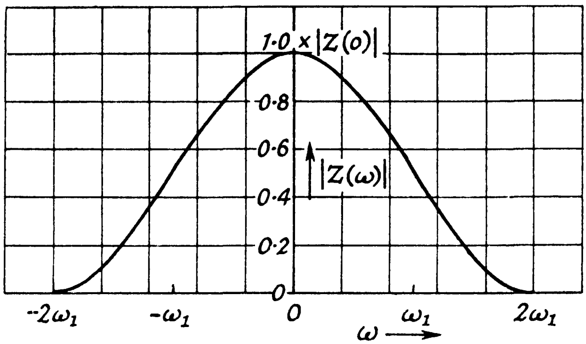
There is, however, one form of intermediate response which is of interest, being the response of an idealised filter with a characteristic

* See also G. A. Campbell (pair 703), Chapter 4, reference 8.

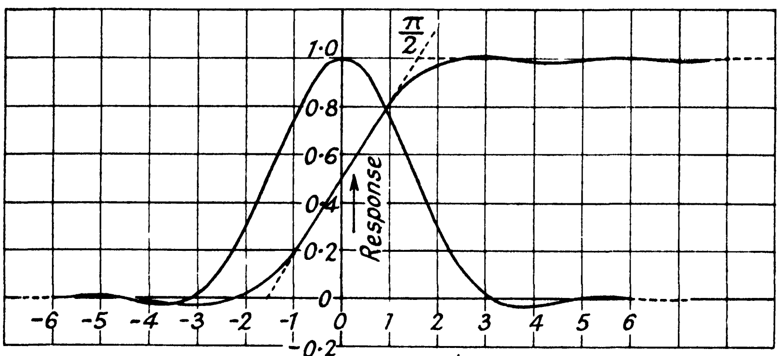
of "cosine-squared" shape. This frequency characteristic, plotted in Fig. 76 (a), has the form of a cosine curve of one complete cycle:

$$|Z(\omega)| = \tfrac{1}{2}|Z(0)| \cdot \left[1 + \cos\frac{\pi\omega}{2\omega_1}\right] \quad . \quad . \quad (238)$$

between the limits $\omega = \pm 2\omega_1$. This has a finite band-width $\pm\omega_1$ between half-amplitude points, but its response has only a minute




(a) The "Cosine Squared" Characteristic.




(b) Response to Infinitesimal Pulse and to Step-wave.

Fig. 76.—The "Cosine Squared" Filter and its Responses.

overswing; also the expression is conveniently simple and the Fourier integral is readily calculable. Applying the infinitesimal pulse $I$, with its uniform cosine spectrum, to such a filter results in the response wave:

$$v(t) = \tfrac{1}{2}\int_{-\omega_1}^{+\omega_1} |Z(0)| \cdot \left[1 + \cos\frac{\pi\omega}{2\omega_1}\right] \cos\omega t \, . \, d\omega \quad . \quad . \quad (239)$$

This is, of course, the Fourier integral of the filter characteristic and may be integrated to give:

$$v(t) \propto \frac{\sin 2\omega_1 t}{(1-4\omega_1^2 t^2/\pi^2)t} \quad . \quad . \quad . \quad . \quad (240)$$

This response is plotted in Fig. 76 (b), together with the response $h(t)$ to a step wave, found by graphical integration. It can be seen that the pulse response is to all intents and purposes finite in duration, lasting a time $t=2\pi/\omega_1$, while the step-wave response reaches its maximum steady-state value in the same time. These are only approximations, but very good and useful ones, being the responses of a filter of *finite* band-width representable by a particularly simple algebraic expression.

These responses are non-physical in that they show finite amplitude before $t=0$, but nevertheless represent an ideal for television, telegraphy, and other similar channels, in that such channels must necessarily have a definite finite band-width allocated to them in order to avoid interference with neighbouring channels.[8] The problem in television is to obtain a response having the most rapid build-up time with least distorting overswing, consistent with the requirement of a finite band-width, provided that the aim of such a system is to reproduce a perfect input signal as closely as is possible.* In this connection the characteristic in Fig. 76 (a), although not absolutely practical, since phase is neglected, represents very nearly the ideal distribution of energy at which to aim.

There is another way of looking at this problem which may be useful. Suppose we wish to transmit a television picture, of which the smallest picture element may be regarded as a rectangular pulse of duration $T_1$, through a channel of finite band-width. Assuming the overall characteristic of the circuits selecting this band to be perfectly flat, as for our original idealised characteristic in Fig. 69, what band-width is the best to use? Fig. 77 (a) shows the rectangular picture element and (b) its frequency spectrum. Also plotted in Fig. 77 (a) is an "equivalent" cosine-squared pulse response wave and on (b) the spectrum of this pulse. It may be seen that this spectrum is very nearly the same as that part of the rectangular pulse spectrum lying between the frequency limits $\omega = \pm 2\pi/T_1$. If,

* Sometimes deliberate introduction of overswing is effected—called plastic distortion—to change the artistic impression of the eventual picture and to correct for the cathode-ray tube spot-size, shape, etc., but this does not affect our present theoretical arguments.

therefore, this band-width is used the picture element received through the channel will closely resemble the "equivalent" cosine-squared pulse response, which is nearly the best that may be done with a finite band-width.

This provides a simple rule for finding the band-width required for a pulse element of given duration. The rule may of course be applied to both band-pass as well as low-pass networks by transferring the mid-band frequency. The curves in Fig. 77 (b) are plotted

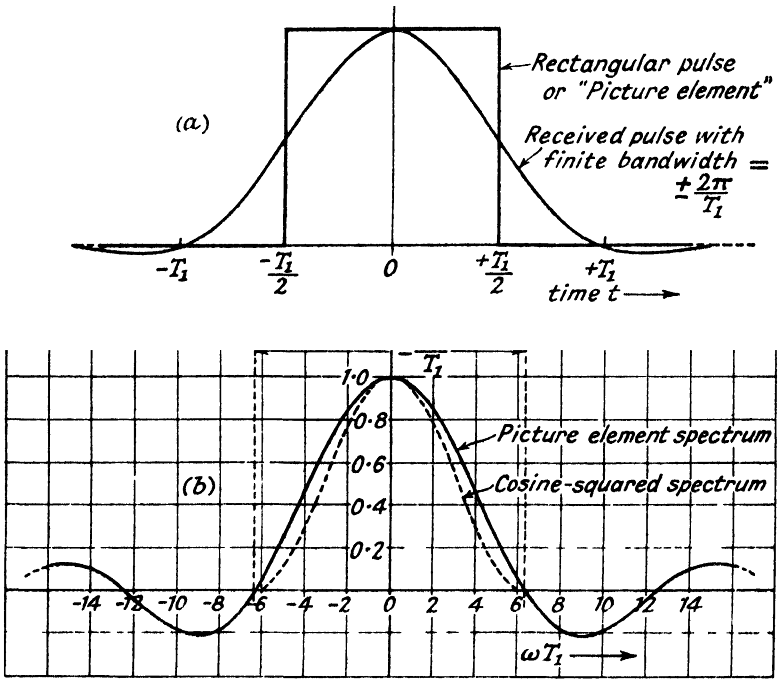



Fig. 77.—Band-width Limitation of a Rectangular "Picture Element."

in conjugate form for this purpose. It is the usual practice, when estimating the time duration $T_1$ of the ultimate "picture element" in a television signal, to consider the picture to consist of a chequer-board of black and white squares of equal period. In this case the signal will consist of a continuously repeated square-wave of periodic time $2T_1$. It should be noted, in this case, that our estimated band-width is such that it will transmit only the fundamental component of this signal, which will therefore be received as a continuous sinusoidal wave of period $2T_1$. But the received signal in the case

of a single picture element (Fig. 77 (a)) is very nearly a single cosine wave of this same period. Thus our conclusions as regards a suitable band-width for single or for multiple picture elements do not disagree to any great extent.

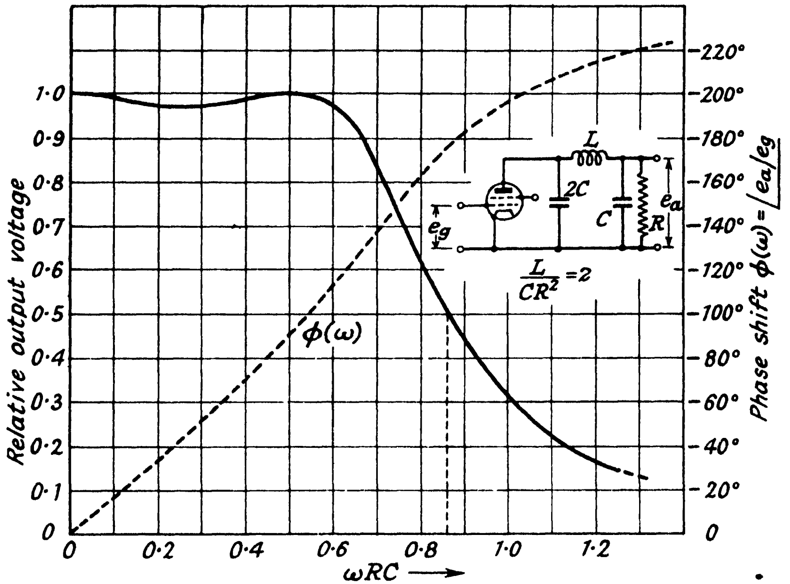
A glance at the spectra of various symmetrical pulse shapes, plotted in Fig. 29 (b), will show that to a first approximation the first main "lobes" of the various spectra are of similar shape. By this expression "first main lobe" is meant that part of a spectrum lying between zero frequency and the first intercept with the frequency axis. Thus the argument which has been applied to the rectangular pulse and its distortion due to band-width limitation may be applied to these other pulse shapes and, indeed, may be extended to almost any other pulse type of signal which has a finite duration. If such signals be passed through filters which accept only the main lobe of their spectra, or even less, the response signals will all be substantially of the same shape, closely resembling the cosine-squared shape of pulse.

If the band-width of a transmitted pulse or signal element be limited by some type of practical filter there will be sufficient phase distortion and delay introduced to remove all traces of signal before the time of receipt of the element, which appear in our theoretical calculations. However, practical response transients may be produced which closely resemble the cosine-squared waveform and which provide an interesting correlation with the idealised responses. For example, the amplifier circuit shown in Fig. 78 (a), a simple constant-$K$ filter $\pi$-section, current-driven by a tetrode,* is frequently used as a shaping network for television signals, while its dual, a $T$-section, voltage-driven by a triode, may also be used [8] (see Chapter 3, Sec. 28). The response $h(t)$ of this circuit to an applied step-wave grid voltage is plotted in Fig. 78 (b), curve (1). Suppose we assume that this filter "cuts-off" at the frequency given by $\omega CR = 0{\cdot}86$, being the frequency at which the amplitude characteristic has fallen to half amplitude, and that this "cut-off" frequency just accepts the first main lobe of the spectrum of a rectangular signal element of duration $T_1$. Fig. 77 (b) shows this spectrum, its first lobe occupying the frequency range $\omega = 2\pi/T_1$. In this case the duration $T_1$ must have the value $T_1 = 2\pi RC/0{\cdot}86 = 7{\cdot}3RC$.

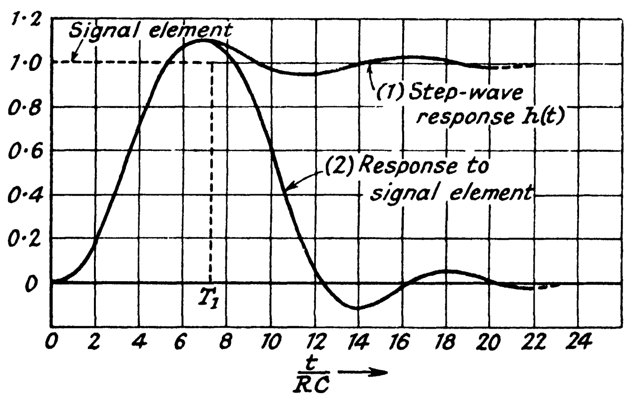
The response of the filter section to a signal element of this duration is also plotted on Fig. 78 (b), curve (2), obtained by

* Note this circuit is similar to that in Fig. 57, but has a full shunt input capacity.

adding a second displaced step-wave response $-h(t-T_1)$ to the first, $h(t)$. This response may be compared with the idealised




(a) Frequency Characteristics of Television Video Amplifier—Constant-$K$ Section with Full Shunt Input Capacity, Driven by a Tetrode.




(b) Response of the Amplifier to a Step-wave and to Rectangular Signal Element.

Fig. 78.—Practical Filter Response Closely Resembling the Idealised "Cosine Squared" Response.

cosine-squared response of Fig. 77 (a), which gives an idea of the error that would have resulted had we taken an idealised filter of

band-width $\omega = 0{\cdot}86/RC$ and avoided the solution of the differential equation necessary for the accurate determination of this network's response. It can be seen from this figure that the general effect of the phase-shift characteristic, which is essentially related to the amplitude characteristic, is (a) to delay the main body of the pulse and (b) to remove the overswings from the leading side and increase them on the trailing side, thereby introducing the asymmetry which is unavoidable in physical networks. This has previously been discussed in Sec. 39 and illustrated by Fig. 60.

In conclusion, the reader is reminded that all the responses we have dealt with here have been plotted as envelope signals, as though they are responses of low-pass filters, but they may be regarded as envelopes of carrier waves, frequency $\omega_0$, if the conjugate network characteristics be transferred in frequency to centre around the frequency $\omega_0$.

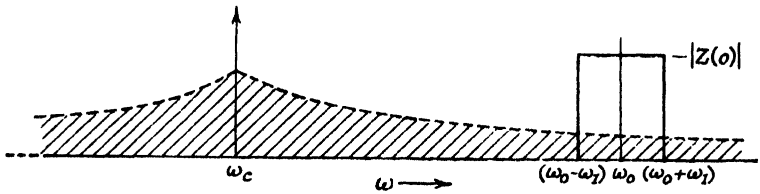
## 51. Response to interfering pulse tuned outside the pass-band

So far we have dealt only with the response of idealised low-pass filters to envelope waves, and with the response of the equivalent symmetrical band-pass filters to modulated carrier waves tuned to the mid-band frequency. In a similar way the idealised characteristics may be used for simple calculation of the interference signal set up in a channel by a signal tuned outside the pass-band.

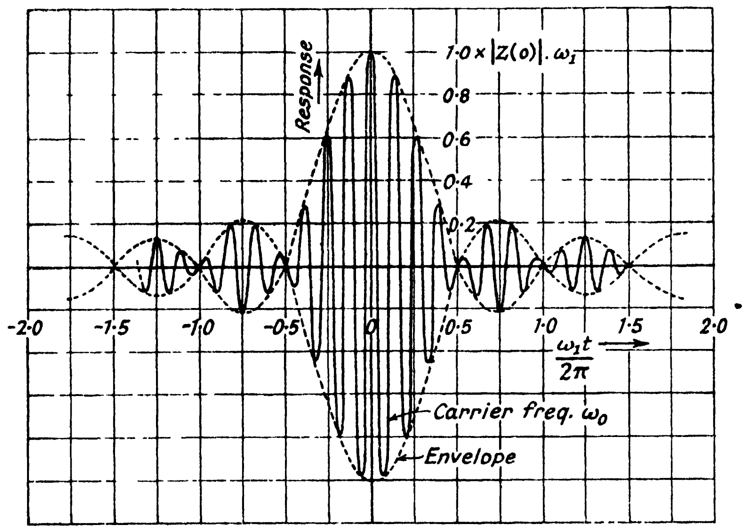
The exact form of this unwanted signal depends on (a) the spectrum of the interfering signal, and (b) the frequency characteristics of the channel. Obviously all kinds of interfering signal may arise in practice and it may happen that the channel is tuned to a maximum of the energy spectrum, thereby receiving a maximum interference, or to a minimum, or to anywhere in between. Once again it is necessary only to consider a single signal element, either a step or a narrow pulse, from which the effect of any interfering signal may be calculated, if necessary, by the Superposition Theorem. This is rarely needed, however, and a broad idea of the interference may be obtained by considering a single element of the interfering signal and by the use of an idealised frequency characteristic.

The simplest case is the rectangular shape of idealised characteristic, and in Fig. 79 (a) such a characteristic is shown tuned to the mid-band frequency $\omega_0$ and covering a part of the spectrum of an interfering modulated carrier wave, frequency $\omega_c$. The spectrum shown here corresponds to a step envelope, but if the frequency separation $(\omega_0 - \omega_c)$ is very great compared with the filter band-width,

the spectrum accepted by this filter may be considered uniform over the band. This assumed uniformity would be exact had we used a narrow pulse for our interfering signal element.




(a) Interference Signal Tuned Outside a Channel.




(b) Interference Response Signal.

Fig. 79.—Response of an Idealised Band-pass Filter to an Interfering Signal Element Tuned Outside the Band.

Addition of all the terms in this uniform continuous spectrum, within the limits of the idealised pass-band, gives the response signal:

$$\text{Interfering signal} \propto |Z(0)| \,.\int_{\omega_0-\omega_1}^{\omega_0+\omega_1} \cos \omega t \,.\, d\omega$$

$$\propto |Z(0)| \left[\frac{\sin \omega_1 t}{t}\right] \cos \omega_0 t \quad . \quad . \quad . \quad . \quad . \quad (241)$$

The absolute magnitude of this response depends, naturally, on the magnitude of the interfering signal element. The form of this interfering signal is shown in Fig. 79 (b); it is a modulated carrier

wave, cos $\omega_0 t$, the frequency $\omega_0$ being the filter mid-band frequency and *not* the interfering signal carrier frequency. The envelope is of the form $(\sin \omega_1 t)/t$ and depends only on the shape of the filter characteristic; it does not depend on the form of the interfering signal provided that this is sufficiently sharp, so that its spectrum uniformly fills the filter band with energy.

The equation 241 may be rewritten:

$$\text{Interfering signal} \propto |Z(0)|\omega_1 \left[\frac{\sin \omega_1 t}{\omega_1 t}\right] \cos \omega_0 t \quad . \quad (242)$$

But at the time $t=0$, $(\sin \omega_1 t)/\omega_1 t = 1$, so that the peak amplitude of the interference signal, which occurs at this time, is proportional to $|Z(0)| \,.\, \omega_1$. This product is the *area* of the filter characteristic, a fact that becomes clear when we remember that the peak amplitude of the interference signal, at $t=0$, is the point at which all the cosine components of the continuous spectrum, passed by the filter, come into phase. With this rectangular form of idealized filter characteristic these components are unaltered in amplitude, but this is not true with other types of filter, such as we have considered in the preceding sections. Another point to be borne in mind is that by ignoring the phase-shift characteristic, which determines so essential a part of practical filter behaviour, the interference wave that has been calculated is somewhat greater in amplitude than it would be in practice. The effect of non-uniformity in the phase-shift characteristic must be to prevent the cosine terms in the interference signal from coming into phase simultaneously at $t=0$ and so to reduce the signal's peak amplitude and to introduce asymmetry in its waveform.

If we confine our attentions here to the idealised filter characteristics and ignore the effects of phase non-uniformity, the form of the interference pulse resulting from filling the band uniformly with energy is readily determined. Its spectrum will be identical with the idealised characteristic $|Z(\omega)|$, whatever the form of this characteristic—rectangular, trapezium, triangular, or the tailing forms in Fig. 75 (a), the cosine-squared form, or any other. If these idealised characteristics are symmetrical about a mid-band frequency $\omega_0$ the resulting interference wave is a carrier wave $\omega_0$ modulated in amplitude by an envelope the spectrum of which has the form $|Z(\omega)|$. This envelope has the same shape whatever the mid-band frequency $\omega_0$, even if this be reduced to zero, giving the equivalent low-pass filter response.

The peak amplitude of the resulting interference wave, occurring at the time $t=0$, is given by the sum of all the components' amplitudes passed by the idealised filter characteristic:

$$\text{Peak amplitude of interference} \propto \int_{-\infty}^{+\infty} |Z(\omega)| d\omega \quad . \quad . \quad (243)$$

which is the area of the characteristic.

If the interfering signal is tuned fairly near to one edge of the channel pass-band, the simple method of treating the problem using an idealized characteristic yields results of doubtful accuracy. In such cases the band will not be filled with a uniform spectrum of energy, but the form of this spectrum may be asymmetric to some extent, depending on the waveform and relative tuning frequency of the interfering signal.

## REFERENCES

1. K. KÜPFMULLER, "On the Relations between Frequency Characteristics and Transient Response in Linear Systems," *Elect. Nacht. Tech.*, Vol. 5, No. 1, p. 18, 1928.
2. E. JAHNKE and F. EMDE, "Tables of Functions," B. G. Teubner, Berlin, 1933.
3. J. B. DALE, "Mathematical Tables," Edward Arnold, London, 1903.
4. D. A. BELL, "Theory of Ideal Filters," *Wireless Engineer*, July 1943.
5. F. ROBERTS and J. SIMMONDS, "Some Properties of a Special Type of Electrical Pulse," *Phil. Mag.*, Vol. 34, p. 822, Dec. 1943.
6. F. ROBERTS and J. SIMMONDS, "Further Properties of Recurrent Exponential and Probability Function Pulse Waveforms," *Phil. Mag.*, Vol. 35, p. 459, July 1944.
7. F. ROBERTS and J. SIMMONDS, "Physical Realizability of Pulse Waveforms of Probability Function Type," *Phil. Mag.*, Vol. 35, p. 778, Nov. 1944.
8. H. NYQUIST, "Certain Factors affecting Telegraph Speed," *Bell Syst. Tech. Jour.*, Vol. 3, p. 324, April 1924.

The following are also of interest:

(A) E. GUILLEMIN, "Communication Networks," Vol. II, p. 477, John Wiley, New York, 1935.
(B) W. P. MASON, "Electric Wave Filters employing Quartz Crystals as Elements," *Bell Syst. Tech. Jour.*, Vol. 13, July 1934, p. 405 (see Fig. 22).
(C) H. E. KALLMAN, R. E. SPENCER, and C. P. SINGER, "Transient Response," *Proc. I. R. E.*, Vol. 33, No. 3, March, 1945.
(D) H. W. BODE and R. L. DIETZOLD, "Ideal Wave Filters," *Bell Syst. Tech. Jour.*, Vol. 14, p. 215, April 1935 (practical filters approaching the ideal flat-topped characteristics).
(E) J. CARSON and O. ZOBEL, "Transient Oscillations in Electric Wave-filters," *Bell Syst. Tech. Jour.*, Vol. 2, p. 1, 1923.

# CHAPTER 6

# Characteristics of Multi-stage Amplifiers

## 52. A chain of amplifier stages—overall frequency characteristics

In the last chapter we have dealt only with the *overall* idealised characteristics of a system and have included one practical example of a single-stage amplifier (Fig. 78) the transient characteristics of which approach closely the ideal "cosine-squared" form. However, it is often required to produce certain desirable forms of response characteristic with a number of stages of valve amplification—for instance in television and radar channels in which a high amplification is needed. In this connection it is interesting to study, from the idealised response point of view, the relations between the characteristics of a single and a multi-stage amplifier.

Amplifier valves may be classified broadly into two types, behaving as voltage generators or as current generators. Thus the triode type may be regarded as a voltage generator having an internal resistance $R_a$ and a *generated E.M.F.* equal to $\mu$ times the applied grid voltage $e_g$, where $\mu$ is the amplification factor; tetrodes, pentodes, and similar types, on the other hand, behave as *current* generators in that the current which flows in the anode lead is equal to $g$ times the applied grid voltage, $e_g$, where $g$ is the mutual conductance, and is independent of the voltage on the anode. Both these are ideal conceptions, since all thermionic valves are non-linear to some degree, $\mu$ and $g$ not being constant. However, in this chapter the non-linear property of valves is ignored, which may be justified by considering only very small voltages to be applied to the grids.

Thus these two types of valve could be taken to represent the voltage and current generators used in Chapter 1, as in the Figs. 1, 3, etc., as well as in later illustrations, and they may be regarded as "inverse" generators when used in connection with dual networks (see Sec. 28). Their equivalent circuits are shown in Fig. 80 (a), with a load impedance $Z$.

It is with tetrode or pentode valve stages that we shall be most concerned, since these valves of the "constant current" type are the most commonly used in multi-stage amplifiers. Fig. 51 shows a

typical amplifier chain of stages together with the responses of the 1st, 2nd, 4th, and 8th stages when a perfect step wave $E$ is applied to the first grid. It has been assumed in the calculation of these responses that the $R_1C_1$ coupling circuit has an infinite time constant so that it contributes nothing towards the distortion of the step wave; furthermore, perfect decoupling of the screens and cathodes is taken for granted, so that the distortion of the step wave, which

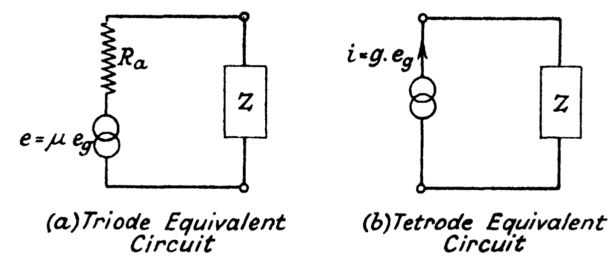



Fig. 80 (a).—Triodes and Tetrodes as Voltage and Current Generators.

is seen to increase with the number of stages, is *high-frequency* distortion* and is due entirely to the characteristics of the anode load, $L$, $C$, and $r$. This distortion takes the form of an increasing build-up time, an increasing bottom curvature, and an increasing overswing as $N$, the number of stages of amplification, is increased. The apparent delay time is not included as a distortion; we say

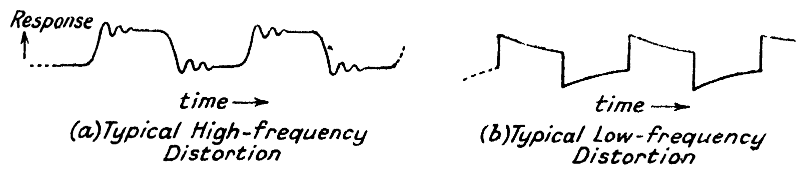



Fig. 80 (b).—Distortion of a Square Wave Typical of (a) the High-frequency and (b) the Low-frequency Ends of an Amplifier Characteristic.

*apparent* delay here because there can be no real time delay in a lumped circuit, but the response remains very small until a certain time has elapsed and then it starts to build up more rapidly. It is convenient to distinguish between this high-frequency distortion and the low-frequency distortion which is usually due to too short or unsuitably valued time-constants in the coupling and decoupling circuits, since these distortions [1, 4, 5] produce entirely different effects

* See reference 7, Chapter 3.

on the step-wave response, examples being shown in Fig. 80 (b). The general effect of low-frequency distortion is a tilting of the whole response waveform without modification of the build-up curve; it is usually possible to eliminate this defect, or at least to reduce it to a negligible value, by suitable adjustment of the decoupling circuit time-constants.[2] The effects of high-frequency distortion are unavoidable to some degree and generally present the greatest difficulty in the design of an amplifier intended to preserve the waveform of extremely sharp impulses or step waves, and this difficulty increases with the number of valve stages required. For this reason, which will be examined in greater detail in the following sections, it is usually desirable to keep the number of stages in such an amplifier down to a minimum and to derive as great an amplification as possible from each stage.

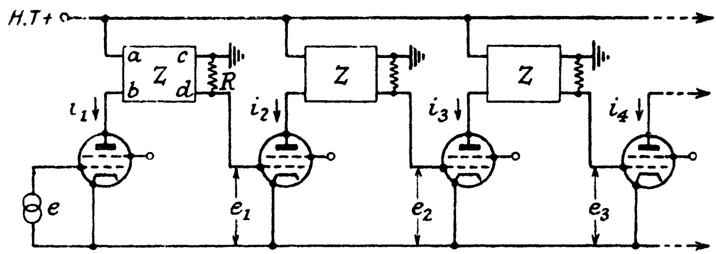



Fig. 81.—A Multi-stage Amplifier Channel (Schematic).

The causes of the increase in high-frequency distortion with $N$, the number of stages, become obvious from a consideration of the frequency characteristics of an $N$-stage amplifier. Fig. 81 shows a schematic circuit of such an amplifier; the decoupling circuits for the screens, the cathode bias, and also the grid coupling circuits are assumed to have infinite time-constants and have been omitted from this diagram.

The anode load, $Z$, shown here as a 4-terminal network (see for example, Fig. 78 (a)), may alternatively be 2-terminal (as in Fig. 51), in which case terminals $a$ and $c$ coincide, as do also terminals $b$ and $d$.

Let $e_g = E \cos \omega t$ be the signal applied to the first valve grid. Then the anode current $i_1$ of this first valve is:

$$i_1 = ge_g = g.E \cos \omega t$$

where $g = \partial i_a / \partial e_g$ the valve mutual conductance, assuming the

perfect conditions mentioned above and also that the valve has an infinite anode impedance. Then the output voltage is $e_1$ where:

$$e_1 = gE \,.\, |Z(\omega)| \cos[\omega t - \phi(\omega)] \qquad . \quad . \quad (244)$$

$|Z(\omega)|$ and $\phi(\omega)$ being the impedance characteristics of the anode load. This voltage $e_1$ is applied to the grid of the second valve, the anode current of which is, similarly, $i_2$ where:

$$i_2 = ge_1 = g^2E \,.\, |Z(\omega)| \cos[\omega t - \phi(\omega)]$$

so that the output voltage of this second valve is $e_2$:

$$e_2 = g^2E \,.\, |Z(\omega)|^2 \cos[\omega t - 2\phi(\omega)] \qquad . \quad . \quad (245)$$

and this voltage is applied to the grid of the third valve, and so on. The anode voltage of the $N^{th}$ stage is then $e_N$:

$$e_N = g^NE \,.\, |Z(\omega)|^N \cos[\omega t - N \,.\, \phi(\omega)] \quad . \quad . \quad (246)$$

Comparing this to the input signal voltage $e = E \cos \omega t$ shows us that the overall characteristics of the $N$-stage amplifier are:

$$\left.\begin{array}{ll} \text{Modulus,} & |Z(\omega)|^N \\ \text{Phase shift,} & N \,.\, \phi(\omega) \end{array}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (247)$$

apart from a constant magnitude $g^N$. This is a special case of the general chain network characteristics discussed in Sec. 27, in which the individual sections of the chain here are unilateral impedances and do not reflect energy back from their output to their input terminals.

Thus the overall amplitude characteristic varies as the power of $N$, while the phase shift varies proportionally to $N$. For instance, consider the frequency characteristics plotted in Fig. 52: at the frequency $\omega = 2{\cdot}5/rC$ the amplitude characteristic has fallen to $\frac{1}{2}$ relative to its value at $\omega = 0$ (which is the ratio $|Z(\omega)|/r$ at this frequency) and the phase shift is 79°. The corresponding characteristics of two similar stages would then be $\frac{1}{4}$ and 158° and of three stages $\frac{1}{8}$ and 237°, and so on. The amplitude characteristic thus becomes narrower and narrower while the phase shift increases its average slope and its departure from linearity, thereby increasing the phase delay time $\phi(\omega)/\omega$ as well as the phase distortion.

## 53. Idealised characteristics of multi-stage amplifiers

Clearly an idealised amplifier stage with the flat-topped characteristic shown in Fig. 69 is a special case. However many such stages there are in a multi-stage amplifier the frequency characteristics remain of the same form and so also must the step-wave

response, which is the idealised sine-integral form calculated in Sec. 46 (Fig. 70). The build-up time and signal distortion of such an idealised amplifier are independent of *N*, the number of stages. Characteristics of such a form cannot be achieved exactly in practice, but it is possible to design filters with extremely flat characteristics [3, 6] suitable for anode loads, and the characteristics of a number of such stages approaches this ideal; their responses can closely approach the sine-integral shape. If such distortion is considered objectionable it is possible to use a multi-stage amplifier with flat-topped characteristics and to follow this with a single signal-shaping stage such as that in Fig. 78.

This is common practice in television receivers, most of the necessary amplification being obtained by a fairly flat-topped I.F. amplifier chain,[23] using either mutually coupled damped tuned-circuit filters [8, 9, 10] or single damped tuned-circuit loads having staggered tuning frequencies*; then a video-frequency amplifier following the second detector is used for shaping the transient response. This system has the advantage that considerable mistuning of the incoming carrier or of the local oscillator circuit is possible before distortion of the signal becomes noticeable—an effect we shall examine at length in Chapter 7.

If the build-up rate of a signal is to be preserved though the number of stages of amplification, *N*, needs to be increased in order to obtain sufficient overall gain, then the band-width of each stage (in any amplifier other than the flat-topped characteristic type) must be widened, depending upon *N*, so that the *overall* band-width remains constant. Since the characteristic of any practical stage cuts-off at a finite rate the choice of the frequency which is considered to represent the "band-width" is quite arbitrary; we have discussed the best way of making this choice in Sec. 46. Furthermore, the *shape* of the overall characteristic changes with *N* in most practical amplifiers, though the arbitrary "band-width" is preserved so that the waveform of the step-wave response varies also, as has been illustrated by the example in Fig. 51.

Let us consider the simplest case of a resistance-loaded amplifier stage: under ideal conditions the response of such an amplifier would be independent of frequency, but the high-frequency characteristics are, in practice, modified by the inevitable anode/earth and grid/earth capacities of the valve and associated wiring. The

* See also the filtering system described by Butterworth (reference 6).

characteristics of a single stage of this type have already been discussed in Sec. 36; Fig. 54 is a diagram of the stage, showing the lumped stray capacity $C$ appearing in shunt with the anode load $R$. The complex impedance of this complete load is given by equation 197, repeated here:

$$Z(j\omega)=\frac{R(1-j\omega CR)}{(1+\omega^2C^2R^2)} \quad . \quad . \quad . \quad . \quad [(197)]$$

from which we may obtain the modulus and phase shift:

$$\left.\begin{aligned}|Z(\omega)|&=\frac{R}{\sqrt{(1+\omega^2C^2R^2)}}\\ \tan\phi(\omega)&=-\omega CR\end{aligned}\right\} \quad . \quad . \quad . \quad (248)$$

Thus, from 247, the overall gain characteristic of $N$ such stages is:

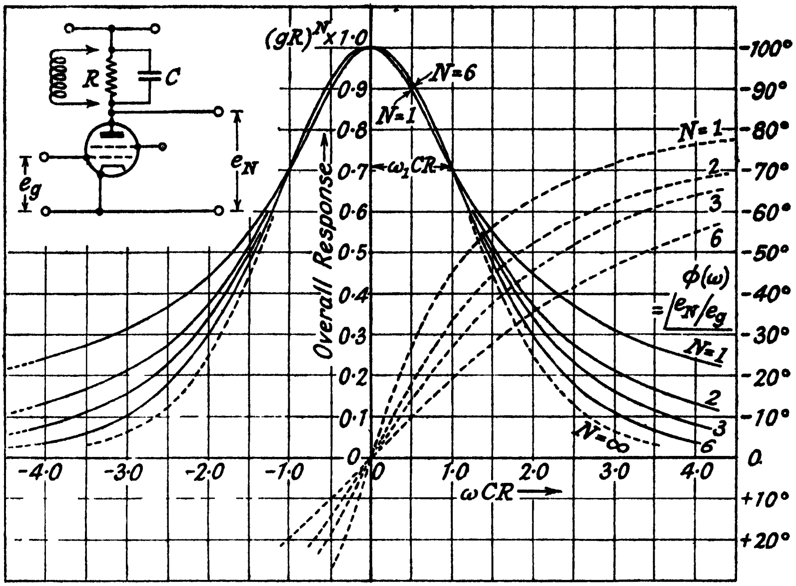
$$\left.\begin{aligned}&\text{Overall modulus} &&=\frac{(gR)^N}{(1+\omega^2C^2R^2)^{N/2}}\\ &\text{Overall phase shift, } N\,.\,\phi(\omega) &&=-N\,.\,\tan^{-1}\omega CR\end{aligned}\right\} \quad . \quad . \quad . \quad (249)$$

These overall characteristics are plotted in Fig. 82 (a) for several values of $N$, in such a way that the band-width at the half-power output point, $|Z(\omega)|=R/\sqrt{2}$, is independent of $N$. This means that every time a stage is added the time-constant $CR$ of each individual stage must be reduced, so that $CR$ is a function of $N$. Let $\omega_1$ represent this band-width, so that, from (249):

$$\left.\begin{aligned}(1+\omega_1{}^2C^2R^2)^{N/2}&=\sqrt{2}\\ \text{or} \qquad CR&=\frac{1}{\omega_1}\,.\,\sqrt{(2^{1/N}-1)}\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (250)$$

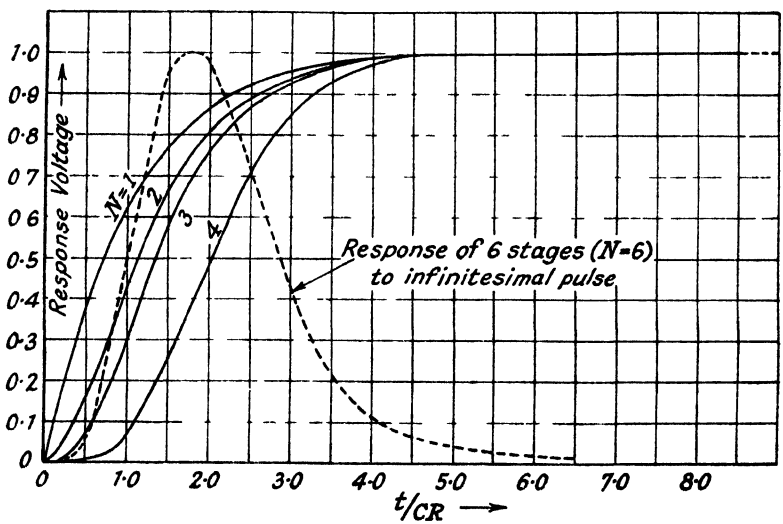
The change in the shape of these overall characteristics, as $N$ is increased but the nominal band-width $\omega_1$ kept constant, may be seen from Fig. 82 (a); the modulus of the characteristics tends to become narrower at low levels, while the phase shift per stage, $\phi(\omega)$, becomes more linear with $\omega$. In the limit $N\rightarrow\infty$ an interesting and useful idealised characteristic is produced, the modulus approaching a "probability" function form,* as shown by the dotted characteristic in the figure, and the phase characteristic becoming linear. This form of characteristic has already been discussed in Sec. 48, though unrelated to practical circuits; here we see that it may be very closely approached by a comparatively simple circuit, the case of six stages, $N=6$, becoming a close approximation.

* A proof of this is given in Appendix.




(a) Conjugate Frequency Characteristics of an $N$-Stage Amplifier.

$CR = \frac{1}{\omega_1}\sqrt{(2^{1/N}-1)}$ where $\omega_1$ = required band-width at $1/\sqrt{2}$ of maximum gain (at zero frequency).




(b) Transient Response to an Applied Step Wave to First Grid.

Fig. 82.—The Characteristics of an $N$-Stage Tetrode Amplifier, with a Resistance Capacity Load.

The equation for this "probability" function characteristic is given by equation 236:

$$|Z(\omega)|=|Z(0)| \,.\, \epsilon^{-(\omega/k\omega_1)^2} \quad . \quad . \quad . \quad [(236)]$$

where in our present example $k=1{\cdot}7$ in order to make $\omega_1$ correspond to the frequency at which $|Z(\omega)|$ has fallen to $1/\sqrt{2}$ of its maximum value, $|Z(0)|$.

The response of this $N$-stage amplifier to an infinitesimally short pulse applied to the first grid may be calculated in the manner outlined in Sec. 43, the Fourier integral method. The overall characteristic (249) may be rewritten in terms of its real and imaginary parts from 197:

$$\text{Overall gain characteristic}=(gR)^N \,.\, \left[\frac{1-j\omega CR}{1+\omega^2C^2R^2}\right]^N$$

$$=(gR)^N \,.\, \frac{1}{(1+j\omega CR)^N} \quad . \quad . \quad . \quad (251)$$

The Fourier integral of 251 is given in Campbell's list, pair 431 (see Reference 8, Chapter 4), which gives the waveform of the pulse response:

$$\text{Pulse response of } N \text{ stages, } \propto \left(\frac{t}{CR}\right)^{N-1} \,.\, \epsilon^{-t/CR} \quad . \quad (252)$$

where $CR$, the time-constant of an individual stage, is given by 250. The actual magnitude of this response is infinitesimal, but by plotting this waveform and integrating graphically the step-wave response, Fig. 82 (b) is obtained. The steady-state response level must be $(gR)^N \,.\, E$.

The usefulness of these response curves, Fig. 82 (a) and (b), may be extended by applying the band-pass/low-pass analogy (see Sec. 33); the frequency characteristics (a), which have been drawn in their conjugate form, with symmetrical amplitude and skew-symmetrical phase-shift components, about zero frequency, may be transferred up to a carrier frequency $\omega_0$ provided that $\omega_0 \gg \omega_1$. They will then represent the characteristics of an amplifier using shunt-tuned-circuit loads, as indicated in the circuit diagram on this Fig. 82 (a), tuned to $\omega_0$. At the same time the transient response curves (b) represent the envelopes of the responses.

The step-wave responses, Fig. 82 (b), show slightly different waveshapes, but all have approximately the same average build-up time. The first response, $N=1$, is the familiar exponential build-up curve for a single resistance-capacity loaded stage; successive

responses for larger numbers of stages tend to become more symmetrical as regards their bottom and top curvature and to have more "virtual delay"—the time measured from $t=0$ to the centre of the build-up curve. In the limit, $N \rightarrow \infty$, the pulse response (252) itself becomes of "probability" function waveform, as we have seen already in Sec. 48. The step-wave response is then the integral of this, which is the curve (2) in Fig. 75 (b).

This type of amplifier circuit has then the interesting property that its frequency characteristic and its pulse response are of identical shape—a practical application of the mathematical idea of a function and its Fourier integral being identical, or what are called "self-reciprocal" Fourier Transforms. This circuit is not a very practical one, inasmuch as it requires an infinite number of stages, but it cannot be said to be *non-physical*, since these "probability" function response curves may be approached as closely as is desired by using a sufficiently large number of stages. The approach is a very rapid one, as may be seen from Fig. 82, and the case of $N=6$ is probably sufficiently close to the "probability" curve for many practical purposes.

## 54. Multi-stage amplifiers with unchanging transient responses

We have, so far, considered two extreme types of idealised characteristics for multi-stage amplifiers, together with their theoretical responses, both of which may be approached by physical circuits: these are (A) the flat-topped characteristic of Fig. 71 (a), and (B) the "probability" function characteristic of Fig. 75 (a), curve (2). The former response transient possesses relatively large overswings, but its shape and build-up time do not change from stage to stage; the latter response transient is of the steady trailing type, approaching the steady-state level asymptotically (see Fig. 82 (b), curve (2)), and although its *shape* does not change from stage to stage, its *time scale* does, so that the actual build-up rate deteriorates with successive stages. Thus the time-scales of the transient responses of successive stages must be compressed for these responses to be absolutely identical. That the shape of the response remains constant in this particular idealised case may be seen from the fact that, theoretically, such a response may be produced by a very large number, $N$, of stages (approaching infinity) of the type shown in Fig. 82, and that therefore the same response will be produced by $(N+1)$ stages. The compression of the time scale, necessary to keep the responses of successive stages identical,

is a constant ratio approximately equal to 1·5 : 1, as may be illustrated by the response curves in this same Fig. 82, which are plotted against a scale of $t/CR$. These show approximately a constant build-up time, provided that $CR$ is changed with the number of stages $N$ according to equation 250:

$$CR \propto \sqrt{(2^{1/N}-1)} \quad . \quad . \quad . \quad . \quad . \qquad [(250)]$$

This fact is illustrated in this figure for $N$ as great as 6, at which value the response is becoming extremely close to the idealised "probability" function form.

Thus if $CR$ is kept constant, as the number of stages $N$ is increased the build-up time will increase by this ratio, $1/\sqrt{(2^{1/N}-1)}$, if measured as the ratio between the first and the $N^{th}$ stage; if measured between any two successive stages the build-up time will increase by the ratio for $N=2$, which is $1/\sqrt{(2^{1/2}-1)}$ or approximately 1·5 : 1. Such response curves that do not change their shape, though their scale may alter, with $N$, are of great interest and importance in television amplifier work, and other forms may exist lying in between these two extreme types which may have better response shapes, that is to say, quicker build-up rates, than the probability function characteristic can give and less overswing than the flat-topped type produces, with intermediate values of time-scale change ratio between 1·5 : 1 and 1 : 1.

Kallman, Spencer, and Singer [7] have investigated this possibility, but have considered sets of idealised step-wave responses rather than idealised frequency characteristics. The validity of the conclusions which may be reached from idealised frequency characteristics is sometimes to be doubted owing to the neglect of the phase shift component, a fact which we have emphasised in the preceding chapter. The alternative way of treating the problem, by taking idealised transient responses and determining the required frequency characteristics, has some advantages, particularly for multi-stage calculations.

In practical cases the phase shift is often of more importance than the modulus of a characteristic; for example, we have shown (Sec. 50) that the response of the amplifier stage in Fig. 78 to a rectangular pulse of a certain duration is very nearly the same as the theoretical response of an idealised characteristic of the cosine-squared form, but it may be seen that the modulus of the characteristic of this amplifier stage is very different from this idealised one (Fig. 76 (a)).

The responses of most practical multi-stage television amplifiers will lie between these two extreme types; if overswings are produced by a single stage they will tend to increase as more stages are added, while build-up time measured between two successive stages will in most instances increase by a ratio lying in between the values 1·5:1 and 1:1.

The ideal step-wave response of an amplifier might possibly be considered to be of the shape shown in Fig. 83 (a), curve (1), rising uniformly from zero to steady-state level during any desired time $T_1$. The response of such an amplifier to an infinitesimally short pulse would be the differential coefficient of this step-wave response—i.e. the rectangular waveform (b). The Fourier integral of this rectangular response then gives the required frequency characteristics of the amplifier (c), curve (1), which have the form:

$$|Z(\omega)| \propto \frac{\sin \omega T_1}{\omega} \quad . \quad . \quad . \quad . \quad . \quad (253)$$

the phase shift being linear with frequency. (The exact form of $|Z(\omega)|$ may be seen from the chart of Fourier Transforms, Fig. 29.) Such a set of frequency characteristics, requiring an infinite band-width, cannot be achieved in practice, but Kallman, Spencer, and Singer have shown that even if such a hypothetical amplifier be considered the step-wave responses of successive stages would be of different forms, curving at the start and finish of the build-up more and more as the number $N$ is increased.

These authors have dealt with another, more practicable, type of step-wave response, which changes its shape only very slightly between successive stages. This response takes the form of a half cycle of a sine wave, as shown in Fig. 83 (a), curve (2). The pulse response of such a stage would therefore be the half cosine wave (b), curve (2), and the required frequency characteristics are given by the Fourier integral of this waveform, (c), (which has also been plotted accurately on the Fourier Transform chart, Fig. 29). Again, this curve represents the required modulus of the characteristics, assuming that the phase shift is linear with frequency. This also requires an infinite band-width, but in this case the curve falls away with frequency very rapidly to zero. Such a half-cycle sine-wave build-up curve may be very closely approached in practice by an amplifier of the type shown in Fig. 51, using damped-tuned-circuit anode loads, but with a lower value of Q than is used in that figure, giving a minute overswing. The best value [7] of Q is 0·6, slightly

less damping than the critical value (Q=0·5). This shape of response waveform is very nearly maintained with a number of successive identical stages, and the time-scale increases by approximately the ratio 1·4:1 between successive stages.

It will be noticed that the step-wave response of the idealised "cosine-squared" filter characteristic, drawn in Fig. 76 (b), is itself very close to this half-cycle sine-wave build-up curve. This curve was considered in Sec. 50 as the best idealised characteristic for a *complete* channel, since it produced the maximum build-up rate and minimum overswing consistent with a finite band-width. Kallman, Spencer, and Singer have shown that it is possible, by using com-

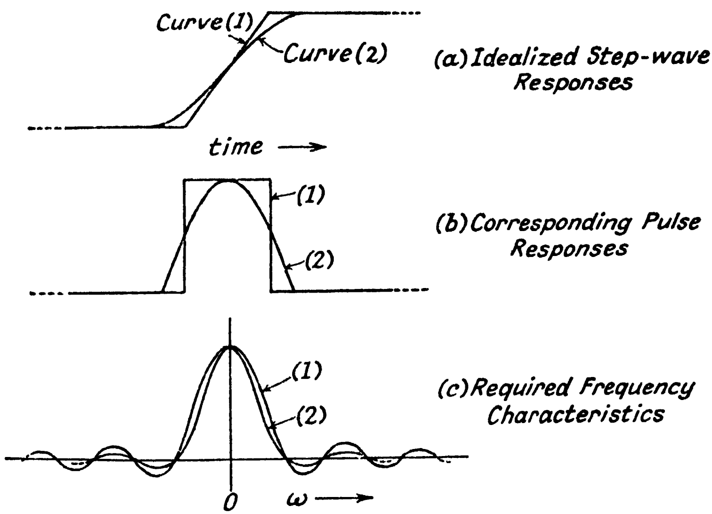



Fig. 83.—Idealised Amplifier Stage Responses.

binations of different anode loads and by careful adjustment of the circuit values, to achieve a practical response which is very nearly perfect in that the build-up rate is a maximum, the overswing is minute, and the responses of successive stages (or groups of stages) remains of constant waveform through the amplifier, their time scales increasing by approximately the ratio 1·4:1 between successive stages. The difference between this waveform and the idealised half-cycle sine-wave build-up curve is extremely small and for theoretical estimations of the behaviour of an amplifier chain this idealised waveform is extremely useful, since the responses of successive stages have very nearly the same waveform but with their

time scales successively increasing by nearly 1·4:1, which may be taken as $\sqrt{2}:1$ for convenience.

To summarise, there are three types of idealised response which are of primary interest. First, the sine-integral build-up curve, Fig. 70, corresponding to the flat-topped characteristic, Fig. 69. This response remains unchanged between successive stages, both as regards its waveform and its time scale. Secondly, at the other extreme, the "probability" function build-up curve, Fig. 75, curve (2), having no overswing. This curve remains of constant waveform, but the time scale increases by the ratio 1·5:1 between successive stages. Thirdly, intermediate between these two, there is the half-cycle sine-wave build-up, Fig. 83 (a), curve (2), which has very nearly the maximum possible rate of build-up and a minimum overswing, and which changes in waveform very little between successive stages though its time scale changes by approximately the ratio $\sqrt{2}:1$.

## 55. The limiting response of an amplifier

The choice of a particular type of network to form the anode load of an amplifier stage, or stages, is obviously influenced by the application which the designer has in mind. At present we are concerned with the use of amplifiers for transient signals, with faithfulness of waveform reproduction as the ultimate aim. This faithfulness will be influenced by the type of anode load and on the number of stages used; these two factors are mutually dependent, since the gain per stage depends on the anode load. In many modern applications we are not free to choose the impedance *magnitudes* of the anode loads (and hence the gain per stage), since the signals being amplified may possess extremely sharp rates of build-up or decay; it is insufficient to preserve only the geometric shape of such build-up or decay curves, but their time-scale must, as near as possible, be unaltered. For example, it may be required to amplify a television signal having signal elements of step waveform rising in 0·1 microsecond from zero to maximum level; considering a build-up curve of half-cycle sine-waveform to be suitable (as discussed in the preceding section), this waveshape could be quite accurately preserved by using a chain of amplifier stages having anode loads of the form shown in Fig. 51 (with a Q value of 0·6). However, the step waveform would deteriorate in build-up *time* as each stage was added, becoming approximately 0·1, 0·14, 0·2, 0·28 . . . microsecond successively, and the resulting television picture would become very poor in definition.

The build-up rate is bound to become poorer from stage to stage in practice, whatever form of anode load is used, and in order to obtain the best response from an amplifier, the overall gain of which is fixed, it is advisable to keep the number of stages down to a minimum by designing each individual stage to give its maximum possible gain consistent with the necessary band-width. The maximum gain per stage is obtained by using an anode load having the maximum impedance, and since a definite band-width is required for a given signal definition, this impedance magnitude is ultimately limited by the inevitable shunt capacities of the valve, appearing between the signal electrodes (anode and grid) and earth.

For example, in the video amplifier shown in Fig. 78 (a), the magnitude of the anode load impedance could be increased by raising the inductance value $L$ and resistance value $R$ and reducing

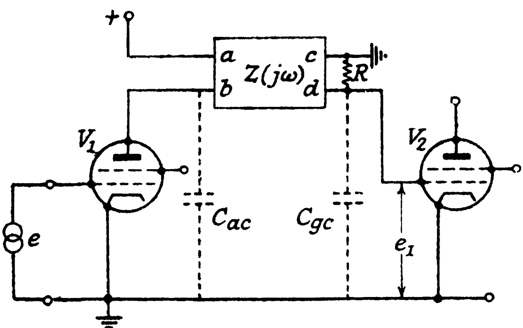



Fig. 84 (a).—Stray Capacities in an Amplifier—Schematic Circuit.

the capacity $C$, all in the same ratio; the horizontal $\omega CR$ scale then remains unchanged, so that the band-width and shape of the characteristics are unaffected,[12] merely the amplifier gain being increased. However, there is a limit to this process, since $C$ cannot be made less than is determined by the valve shunt capacity. In practice with most television amplifiers, whether video or I.F., the load networks [3] are designed with the valve shunt capacities as the basic capacity elements, no extra capacities being added; this gives the maximum value of $R$ for any particular form of load network and hence the maximum stage gain for a given band-width, or vice versa.

Fig. 84 (a) shows these valve capacities as they appear in an amplifier, $C_{ac}$ being the anode to cathode capacity of one valve and $C_{gc}$ the grid to cathode capacity of the next valve, both including stray capacities. They are shown in this schematic circuit as being

separated by some 4-terminal network (as we have used in Figs. 57 and 78 (a)). Sometimes they are directly in parallel, Fig. 84 (b), when the anode load is a 2-terminal network (as in Fig. 52), in which case the terminals *a* and *c* and also *b* and *d* may be considered to coincide.

If the capacities $C_{ac}$ and $C_{cg}$ are separated by some reactive net-

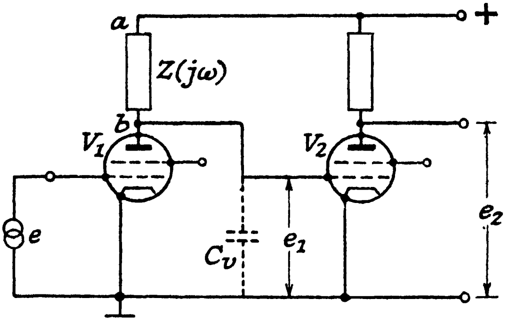



Fig. 84 (b).—Stray Capacity With 2-Terminal Anode Loads—Schematic Circuit.

work, so that the whole effective anode load becomes a 4-terminal network, this load may form a section, or possibly a number of sections, of a filter network. This filter may be a low-pass or a band-pass type. For example, those amplifier loads in Figs. 57 and 78 (a) are constant-*K*, low-pass filter sections; while that in Fig. 87 (b) is a band-pass filter. This combination of amplifier

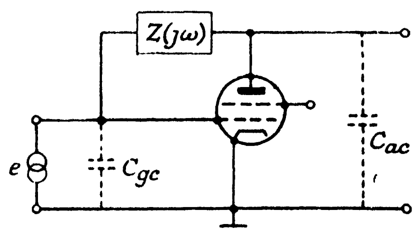



Fig. 84 (c).—Stray Capacities With a Feedback Impedance—Schematic Circuit

valves and filter sections is in common use for obtaining a high and uniform gain over a wide band of frequencies, such as may be required for a television signal. For the design of such amplifiers it is sometimes convenient to use the conventional design data for filter structures, and much published analysis is based on this method. It must be remembered, however, that the conventional

design data are based on iteratively terminated structures, that is to say structures that theoretically have an infinite number of sections. In cases of one or of a few sections terminated by a resistance, such as may be used for the amplifier anode loads, the characteristics will differ from those of the filter prototypes, and this difference can become important when the exact shape of a response signal is in question.

Let us consider now the greatest rate at which an amplifier can respond to a step-wave signal, the limitation depending on the inherent shunt capacities $C_{ac}$ and $C_{gc}$.

The greatest rate at which $e_a$, the anode potential of a single valve, can be made to change is given by removing any anode load, resistive or reactive, leaving the valve to pass all its anode signal current into its own anode/cathode capacity:

$$\left(\frac{de_a}{dt}\right)_{max}=\frac{i}{C_{ac}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (254)$$

If $i$ is made a step wave of current $I$, by making the grid signal voltage $e$ a step wave of amplitude $E$, all this steady anode current will flow into this shunt capacity, giving the uniform rate of change:

$$\left(\frac{de_a}{dt}\right)_{max}=\frac{E\,.\,g}{C_{ac}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (255)$$

If this valve be coupled to a second valve and a 2-terminal anode load be used, then the capacity $C_{gc}$ is added to $C_{ac}$, so that the anode voltage $e_a$ and also the following grid voltage $e_1$ have the rate of change:

$$\left(\frac{de_a}{dt}\right)_{max}=\frac{E\,.\,g}{C_{ac}+C_{gc}}=\frac{E\,.\,g}{C_v} \quad . \quad . \quad . \quad (256)$$

$C_v$ being this total valve capacity $C_{ac}+C_{gc}$.

The ratio $g/C_v$ for any particular valve may be taken as a *figure of merit*, since it determines the maximum rate of response of the valve as an amplifier. Thus for an amplifier to have the best response rate and hence to give the best signal definition, the mutual conductance should be high and the anode and grid shunt capacities low.*

If the two shunt capacities are separated by a 4-terminal network the effective total capacity is not $C_{ac}+C_{gc}$, but is less, and Wheeler [3]

* This requirement has led to the development of special high ratio $g/C_v$ valves; extremely high ratios may be obtained by the use of secondary emission, as in the electron-multiplier type of valve. See reference 14.

has given it as the geometric mean of the two, the valve figure of merit being $g/\sqrt{C_{ac}C_{gc}}$ in this case. This figure is better than that for a 2-terminal load, and it is possible to obtain faster responses by using 4-terminal types of load impedance, though not always having the desired waveform.

Let us consider the maximum possible rate of rise [16] of the anode potential of the $N^{th}$ valve in an amplifier chain using 2-terminal anode loads, as in Fig. 84 (b). If $C_v$ is the total shunt capacity, the maximum rate of rise of the first anode is given by (256) and occurs when there is no load in the anode. From (256):

$$e_a=\frac{E\,.\,g}{C_v}\,.\int dt=\frac{E\,.\,g\,.\,t}{C_v}+K_1 \quad . \quad . \quad . \quad (257)$$

If the circuit is initially at rest, $e_a=0$ at $t=0$ so that the constant $K_1=0$. This steadily changing voltage is passed on to the grid of the second valve, being the voltage $e_1$. Thus the second valve anode potential is:

$$e_2=\int\frac{e_1 g}{C_v}\,.\,dt=E\left(\frac{g}{C_v}\right)^2.\int t\,.\,dt=\frac{E}{2}\left(\frac{gt}{C_v}\right)^2+K_2 \quad . \quad . \quad (258)$$

again the constant $K_2=0$ if $e_2=0$ at $t=0$. This voltage is passed on to the grid of the third stage, and so on. Repetition of this process of integration gives the anode potential of the $N^{th}$ stage, $e_N$:

$$e_N=\int\frac{(e_{N-1})}{C_v}g\,dt=\frac{E}{N!}\left(\frac{g\,.\,t}{C_v}\right)^N \quad . \quad . \quad . \quad (259)$$

In any practical case the anode load impedance $Z(j\omega)$ will be finite, and this anode voltage $e_N$ will then only obtain at the instant $t=0$. After this instant some of the current will start to flow through the anode load and the anode voltage will be less than $e_N$.

Now let us assume that, whatever the form of the anode load, this has a resistive value $R$ at zero frequency, or $|Z(0)|=R$. Then the steady-state amplification of the grid voltage $E$ becomes $gR$ after some interval of time, depending on the initial transient set up by the reactive elements of the load after the sudden application of this grid voltage. The steady-state maximum output of the $N^{th}$ stage is thus:

$$(e_N)_{max}=E\,.\,(gR)^N \quad . \quad . \quad . \quad . \quad (260)$$

after this interval of time. Hence the output voltage $e_N$ given by 259 may be expressed as a fraction of this maximum steady-state response voltage (260):

$$\frac{e_N}{(e_N)_{max}}=\frac{1}{N!}\left(\frac{t}{RC_v}\right)^N \quad . \quad . \quad . \quad . \quad . \quad (261)$$

This is, strictly, accurate only at the instant $t=0$, but it is a useful way of expressing the *maximum* rate of build-up of the potential of the $N^{th}$ valve anode, and is reasonably accurate provided that values of $e_N \ll (e_N)_{max}$ are considered only. The shape of this initial response is plotted in Fig. 85 from equation 261, for $N=1$, 2, 3, 4, 6, 8 and 10. If we consider some small fraction for $e_N/(e_N)_{max}$, say 10 per cent., then we can see from this figure that there is a *virtual* delay time, which increases with the number of stages of amplification, before the output signal $e_N$, of the whole amplifier, has built-up to this fraction of the eventual steady-state output signal $(e_N)_{max}$. Build-up curves of this general form are typical also of passive low-pass filter structures consisting of a ladder of shunt capacities joined by bilateral impedance elements. In such filters

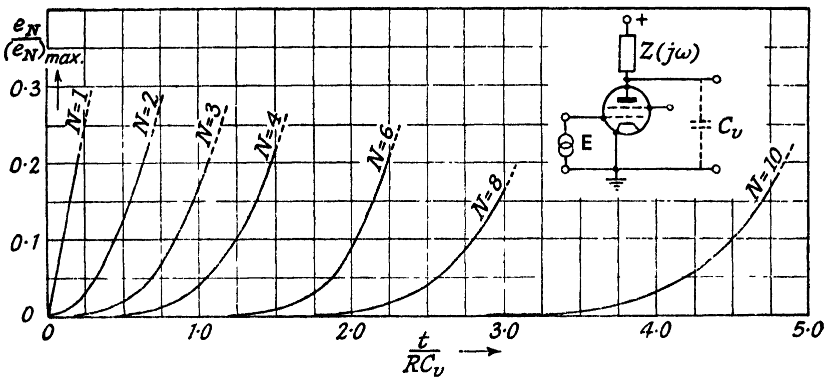



Fig. 85.—Theoretical Maximum Build-up Rates for an $N$-Stage Amplifier.

there is interaction between the various sections, unlike the chain of amplifier stages, but the increase of virtual delay and curvature of the build-up are very similar.[17]

In a practical amplifier the shape of the build-up curves of the various stages depends on the particular form of the anode load, but these curves in Fig. 85 give a theoretical upper limit to the anode voltage build-up of any stage, and show also the minimum possible virtual delay. It is interesting to compare the build-up rate and virtual delay of the response curves of Fig. 51,* for a practical type of amplifier, with these theoretical limiting curves; the virtual delay of a chain of amplifier stages is illustrated even more markedly in a

* Also those of Fig. 82 (b), for which it must be remembered $RC \propto \sqrt{(2^{1/N}-1)}$.

publication by Bedford and Fredendall [15] in which the build-up curves are shown for amplifiers having as many as sixty-four stages. In a case of so many stages the output response signal remains substantially zero for a very long time and then starts to build-up, as though the wavefront has a definite (virtual) velocity through the amplifier, which is due, as we have seen, to the effect of the inherent shunt capacities of the valves.

The simplest form of anode load that we may use is a resistance; in this case $Z(j\omega)$ becomes $R$ in Fig. 84 (b). The presence of this resistance will affect the rate of build-up of the signal, reducing its average value below the limiting value we have calculated. For instance, the response at the anode of the first stage will start to build-up at the rate $E \,.\, g/C_v$, but this rate will reduce to zero exponentially, corresponding to the steady maximum signal $E \,.\, gR$. Fig. 86 illustrates the response curves for a given valve ($g$ and $C_v$ fixed) with different resistive anode loads $R_1$, $R_2$, $R_3$ and $R_4$. The decay curves corresponding to a negative step of $-E$ at the time $t=T$ are also illustrated, the whole set of responses thus corresponding to a rectangular waveform signal of duration $T$, the time $T$ being chosen arbitrarily in this figure.

Now whereas the build-up curve starts at $t=0$ from initial rest conditions, the decay curve which starts at $t=T$ adds on to the continuation of the build-up curve (by the Superposition Theorem, as explained in Sec. 42 and illustrated by Fig. 67 (a)), so that the decay is essentially slower than the build-up, giving an asymmetric response. The amount of asymmetry clearly increases as $T$ is shortened relative to the time-constant $RC_v$. If $RC_v \ll T$ the build-up curve practically reaches its steady-state value before the decay curve starts. We are usually interested in obtaining responses of nearly equal average build-up and decay times, producing nearly symmetric responses to square pulses; an improvement over the plain resistance load may be secured by associating with $R$ certain reactive elements which can have the effect of keeping the rate of change of the anode potential close to the theoretical maximum value $E \,.\, g/C_v$ for a longer time. This is illustrated by the dotted line in Fig. 86. We have chosen the first valve, with its exponential build-up curve, merely for example here, but the same principles apply to any other stage—whatever the shape of the build-up curve, the response to a pulse or picture element of duration $T$ becomes more asymmetrical as $T$ is shortened compared with the time-constant $RC_v$, but the symmetry is improved by preserving the

build-up rate as near as possible to (but never exceeding) the limiting value for any stage, given by equation 261.

We have already discussed various forms of anode load which may be based on the essential resistance $R$ and associated reactive elements. The most important have been illustrated by Figs. 51, 57, and 78, the particular response curves of which have been dealt with in turn. We are not concerned here with the waveforms of these responses, but rather with magnitudes, in order to determine the factors which limit the stage gain and the absolute rate of response. It has been shown above that, in order to amplify a pulse of duration $T$, the basic time-constant $RC_v$ must be kept short; the anode loads mentioned all have a capacity element $C_{ac}$ between the anode and cathode and also $C_{gc}$ between grid and cathode, and

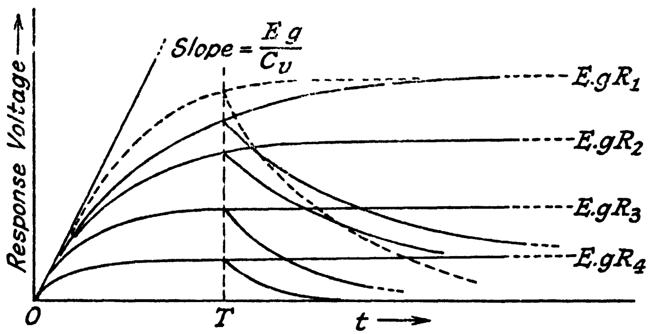



Fig. 86.—Response of a Resistance Loaded Tetrode Stage to an Applied Step Wave. Load Resistance $R$ has the Values $R_1 > R_2 > R_3 > R_4$.

so these may be comprised of the valve shunt capacities, with no additional condensers unless these are essential for balancing up $C_{ac}$ and $C_{gc}$ to the right ratio.

With $C_v$ (or $C_{ac}$ and $C_{gc}$) fixed, $R$ must have a sufficiently low value, the required value depending on the form of anode load circuit used (which decides the build-up waveform) and the time $T$.

There is clearly a lower limit to the duration $T$ of a pulse or picture element beyond which the amplification of a given valve is less than unity, whatever the anode load. This occurs when the shunt capacity of the valve, charging or discharging at the rate $E \, . \, g/C_v$, changes the anode potential during the time $T$ by an amount just equal to the change in grid potential $E$.

In making the above examination of the practical limitations on the response of an amplifier we have considered this problem from

the point of view of video signals, taking the response to a step wave as the criterion of signal reproduction quality. We may, however, extend the conclusions we have reached to band-pass amplifiers,for modulated carrier signals, if we consider the band-pass equivalent circuits of the video-frequency loads (see Sec. 41). In this case the zero-frequency impedance of the load (the resistance $R$) becomes the impedance of a band-pass load at the mid-band carrier frequency, and is itself a resistance in many types of circuit used in practice. The greatest stage gain together with the fastest response of the envelope of the carrier wave is obtained, in the band-pass case, by building the anode filter load to have the highest impedance at mid-band and adding no further capacities to the existing valve self-

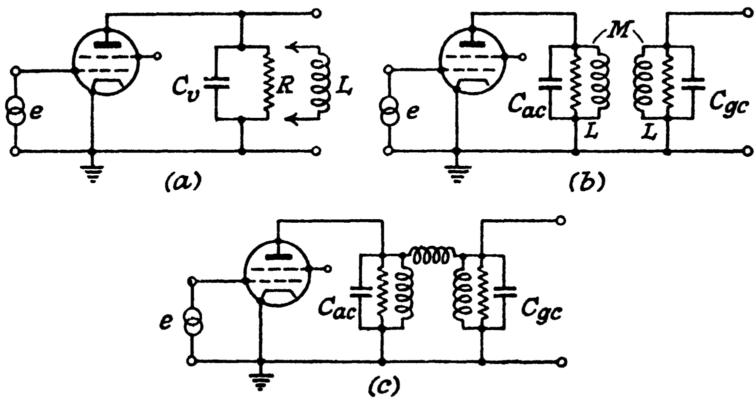



Fig. 87.—Schematic Amplifier Stages (a) with Band-pass/Low-pass Equivalence, (b) and (c) with no Low-pass Equivalence.

capacities; the envelope rates of rise are then limited by the factors that we have determined, illustrated in Fig. 85.

These arguments do not, theoretically speaking, apply to band-pass filters of non-symmetrical structure which have asymmetric frequency characteristics, since such filters do not have low-pass (i.e. video-frequency) equivalents, as was seen in Sec. 33. However, in most practical circuits (for example in television I.F. filters) the band-pass characteristic asymmetry is very slight and the general conclusions are applicable. In cases of serious side-band asymmetry the rates of build-up of the envelope signal response are somewhat reduced, a phenomenon which is examined in detail in Chapter 7. Fig. 87 illustrates these points: the amplifier (a) is essentially a resistance-loaded stage with the shunt capacity $C_v$ (shown schematically)

but its behaviour may be simulated at carrier frequency, with regard to its response to the envelope of the applied wave, by tuning with an inductance $L$ to the carrier frequency. On the other hand, the band-pass amplifier (b) has no "low-pass analogue" of equivalent behaviour, since the circuit is of non-symmetrical structure, as may be seen from the equivalent circuit (c), one arm of which is untuned.

## 56. Constancy of the product: Band-width × Gain

In connection with the problem of determining the ultimate definition of which an amplifier chain is capable, the steady-state approach presents some interesting and useful ideas. In the following notes we shall again treat the low-pass amplifier case, but the application of the results to band-pass amplifiers may be inferred.

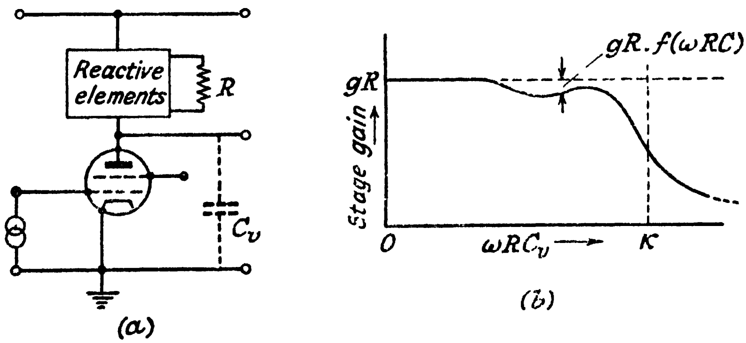



Fig. 88.—Schematic Amplifier Stage with its Gain Characteristics.

The first point is that if an amplifier has been designed to give the maximum gain consistent with a certain band-width, this band-width may only be increased at the expense of gain; the product of band-width and gain is constant for a given type of valve and load circuit structure. This statement needs some elaboration, so let us refer to Fig. 88. This shows, schematically, an amplifier stage (a) loaded with a resistance and certain reactive elements having the transfer characteristics (b). These reactive elements must include the shunt capacities of the valve anode and grid to ground, $C_{ac}$ and $C_{gc}$, and since the stage is to give its maximum possible gain these valve capacities represent the entire shunt capacities of the anode load, none extra being added. It is to be assumed that the values of the basic time-constant $RC_v$ and of the associated reactive

elements have been chosen to give tolerable signal build-up time and waveform distortion. The characteristics, Fig. 88 (b), have been plotted against a scale of $\omega RC$ as "universal" characteristics* (see Sec. 28); practical examples of such curves are to be found in Figs. 33 (c), 52, 57, and 78 (a), but as an alternative to this reference time-constant $RC_v$ a reference frequency $\omega_0$ may be chosen and characteristics plotted in terms of $\omega/\omega_0$, as in Figs. 44 and 46, in which case $1/\omega_0$ may be expressed as a multiple of $RC_v$.

For example, take the amplifier characteristics of Fig. 78 (a). If this 4-terminal load is to be designed to have the maximum possible impedance $|Z(\omega)|$ in order to produce the greatest stage gain, the first shunt capacity element $2C$ must be made equal to the valve anode-to-cathode capacity $C_{ac}$, while the second shunt capacity element $C$ must correspond to the grid-to-cathode capacity $C_{gc}$ of the next valve. Two difficulties may arise in this: (a) the valve shunt capacities may not have the required ratio of 2:1 and (b) there may be no second valve coupled to the first. In either case the shunt elements must be padded up to the necessary values by additional condensers.

In this particular example, Fig. 78 (a), let us choose as the reference frequency, $\omega_0$, the frequency at which the gain has dropped to half amplitude. Then:

$$\omega_0 RC = 0{\cdot}86 = \kappa$$

This frequency, $\omega_0$, may be taken to represent the nominal bandwidth of the amplifier, but it is a purely arbitrary choice. The amplifier gain at zero frequency is given by $gR$, but if the resistance $R$ be reduced to $R/n$, where $n$ is any constant, this gain falls to $gR/n$. At the same time the reference frequency $\omega_0$ must increase to $n\omega_0$ in order to keep the value of $\kappa$ constant:

$$n\omega_0 \,.\, \frac{R}{n} \,.\, C = \kappa \qquad \ldots\ldots \qquad (262)$$

at which point the gain characteristic must still fall to half amplitude. At the same time the inductance $L$ must be reduced to keep the value of $L/CR^2 = 2$, a necessary condition for preserving the shape of these characteristics; it will be remembered that $\sqrt{(L/CR^2)}$ represents the effective Q of the circuit (Sec. 28). If a more complex type of

* The use of $C_v(=C_{ac}+C_{gc})$ really assumes a 2-terminal load, but a similar time-constant may be used for plotting 4-terminal load characteristics; either $RC_{ac}$ or $RC_{gc}$ will serve, for example, since such a constant is merely a reference figure for the frequency scale.

load be considered, containing a number of meshes, the various elements have to be adjusted in the same way so that the Q values of every mesh are unaltered with the value of $\omega_0$.

We may write the product of gain and nominal band-width thus:

$$gR\,.\,\omega_0=\omega_0 RC\,.\,\frac{g}{C}=\kappa\left(\frac{g}{C}\right) \quad . \quad . \quad . \quad (263)$$

which is a constant.

This product is proportional to the ratio $g/C$, which itself depends on the "figure of merit" for the valve, $g/C_v$.

Fig. 88 (b) illustrates a typical gain/frequency characteristic of an amplifier stage, which has the gain $gR$ at $\omega=0$ and departs from this value by varying amounts at higher frequencies. We may express these departures from the steady value $gR$, representing amplitude distortion, as a function of $\omega RC$, so that the gain characteristic may be written:

$$\text{Stage gain characteristic}=gR[1-f(\omega RC)] \quad . \quad . \quad . \quad (264)$$

where $f(\omega RC)$ must have small values compared with unity in order that the variation in gain, at any frequency contained in the spectrum of the applied signal, may be small. The overall gain of $N$ such stages is:

$$\text{Gain of } N \text{ stages}=(gR)^N[1-f(\omega RC)]^N \quad . \quad (265)$$

as we have already shown—equation 247, Sec. 52.

Expanding this expression by the Binomial theorem:

$$\text{Gain of } N \text{ stages}=(gR)^N\left[1-N\,.\,f(\omega RC)+\frac{N(N-1)}{2!}\,.\,f(\omega RC)^2+\,.\,.\right]$$

$$=\text{G}[1-N\,.\,f(\omega RC)] \text{ approximately} \quad . \quad . \quad (266)$$

where $\text{G}=(gR)^N$, the overall steady-state gain. Thus the overall distortion, $N\,.\,f(\omega RC)$, is $N$ times the distortion of one stage.

As regards phase distortion, if $\phi(\omega)$ is the phase shift of one stage, the total phase shift of $N$ stages is $N\,.\,\phi(\omega)$, as we have shown in Sec. 52, equation 247, so that the phase distortion also increases proportionally to $N$.

Negative feedback and its effects on the characteristics of a multi-stage amplifier have not yet been mentioned. It is possible, by the application of negative feedback, to reduce these departures in the gain characteristic from the steady value $gR$, and so effectively to widen the useful band over which the gain is reasonably uniform.

However, the gain will be reduced by the negative feedback at the same time. Whether the product of band-width and gain is any greater than that obtained without the application of feedback is doubtful.

The general problem of feedback in a multi-stage amplifier, which takes place between various stages in numerous ways, is a very difficult one,[25] and no general solution is to be found.* A more readily calculable result is obtained by confining the feedback to individual stages, applying it between the anode and grid of any valve. This feedback may be of several kinds: thus it may be current or voltage feedback, and may vary with frequency or not. Fig. 84 (c) shows, schematically, an interesting case of an amplifier stage in which energy is fed back from the anode circuit to the grid by the feedback impedance $Z(j\omega)$, so that such a stage may be considered to be bilateral in that a signal may pass one way through the valve and in the reverse direction through the impedance $Z(j\omega)$; the forward and return signal paths are thus different, but as regards transmission of a signal between the input terminals and the output terminals, this network may behave like a bilateral 4-terminal network.

The stray capacities $C_{gc}$ and $C_{ac}$ appear across these input and output terminals, and these capacities are connected by $Z(j\omega)$ (between their live terminals, the others being earthed), which may be made reactive. Thus there are all the essentials for the 4-terminal network to behave like a filter section,[3] and adjustment of the various circuit constants—the valve parameters, the value of $Z(j\omega)$ and the form of the feedback coupling—can set the cut-off frequency to the required value. Circuits of this type may then be arranged to have characteristics similar to those already discussed in which normal filter sections with resistance terminations are used as anode loads. It is not possible to give general conclusions as to which form of amplifier is best, with or without feedback, from the point of view of amplifying sharp pulses and transients, and a detailed analysis would be too lengthy to be entered into here.[2, 3]

## 57. The optimum number of amplifier stages

Suppose it is required to design an amplifier with an overall gain ratio of G and a certain degree of distortion. The problem arises:

* See, for example, W. H. Bode's book, "Network Analysis and Feedback Amplifier Design," Van Nostrand, 1945.

how many stages, $N$, should be used? We have seen in the preceding section that the product of gain and band-width is a constant, but the "band-width" is an arbitrary quantity. The gain characteristic of one stage will vary from the steady gain $gR$ at zero frequency (or the mid-band frequency in the case of a band-pass amplifier) as in Fig. 88, the greatest departures from this value occurring at the high frequency end in most practical amplifiers. Thus it might seem that the best results would be obtained by designing very wide band stages, so that the applied signal could be accommodated on the relatively flat low-frequency end of the characteristic. Such a band-width might, however, necessitate an extremely low stage gain and hence a large number of stages $N$, or even a stage gain less than unity. With a very large number $N$ the slight variations in the gain characteristic and, even more important, the phase-shift characteristic multiply up and can eventually give

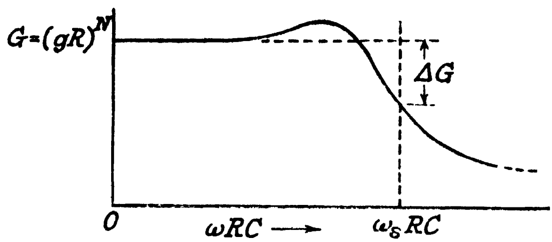



Fig. 89.—Gain Characteristic of $N$-Stage Amplifier.

intolerable distortion. On the other hand, the stage gain can be made as high as possible, and $N$ kept to a minimum, by reducing the band-width until the signal spectrum is seriously distorted by the high-frequency end of the gain characteristic.

The way out of this dilemma is to be found by considering what *overall* distortion of the whole amplifier chain is tolerable and then apportioning this distortion equally amongst a number, $N$, of stages (assuming the stages to be identical), eventually settling on the optimum number by a series of judicious guesses. Thus suppose it is decided, as a measure of distortion, that the overall gain $G$ may drop to $(G-\Delta G)$ at a certain upper limiting frequency $\omega_s$ (see Fig. 89), then the following procedure may be adopted:

(a) Choose some reasonable number of stages $N$ and divide the overall drop in gain, $\Delta G$, into $N$ equal parts (see equation 266—an approximation).

(b) From the characteristic curves of a single stage find the value of $\omega RC$ at which this drop* in gain $\Delta G/N$ occurs. Since this point is to be at the frequency $\omega = \omega_s$ and also since $C$ is fixed by the valve capacities, this decides the value of $R$.

(c) See whether this value of $R$ gives the required overall amplifier gain $G$ $[=(gR)^N]$.

If this is unsatisfactory a new value of $N$ must be chosen, and so on, until the required gain is attained. In practice the optimum value of $N$ is very quickly found, but in any case it is not at all critical except when anode loads are used having very sharp cut-off rates. Exact calculation of the optimum number of stages is possible only when the analytical form of the gain characteristic of a single stage is known, but the practical value of such calculation is limited.[22] The general method of this calculation should be indicated:

The gain of one stage at zero frequency is $gR$ and at any other frequency it is $g|Z(\omega)|$, where $|Z(\omega)|$ is the impedance of the anode load. It is convenient, as we have seen, to plot these gain characteristics in terms of $\omega RC$ rather than $\omega$, and on this basis the drop in gain at any frequency may be expressed as $gR - g|Z(\omega RC)|$; but this drop in gain has previously been written as $gR . f(\omega RC)$ (equation 264 and Fig. 88). Hence we may write:

Gain characteristic of one stage $= gR[1 - f(\omega RC)]$ . . . . [(264])

But at the frequency $\omega_s$ the gain $G$ of $N$ stages is required to be $G(1-\Delta)$ [where $G=(gR)^N$]; thus:

$$G[1-f(\omega_s RC)]^N = G(1-\Delta)$$

In the expression $f(\omega_s RC)$ we may write $(\omega_s RC)$ as $\left(\omega_s \,.\, gR \,.\, \dfrac{C}{g}\right)$, so that:

$$\left[1-f\left(\omega_s \,.\, G^{1/N} \,.\, \frac{C}{g}\right)\right] = (1-\Delta)^{1/N} \qquad . \quad . \quad (267)$$

The value of $N$ to give the maximum value of $\omega_s$ may then be found by differentiating with respect to $N$ and equating to zero, in the usual way.

It must also be remembered that the phase-shift characteristic will probably determine the signal distortion to a great extent,

* We may of course work on the phase-shift characteristic instead, choosing a frequency $\omega_s$ at which this just gives tolerable distortion.

particularly if $N$ is large. In this case the value of $\omega_s RC$, corresponding to the highest wanted frequency in the signal spectrum, must be chosen on a basis of the tolerable phase error—the amount by which the phase-shift characteristic lying between $\omega=0$ and $\omega_s$ departs from the ideal linear (uniform delay) characteristic.

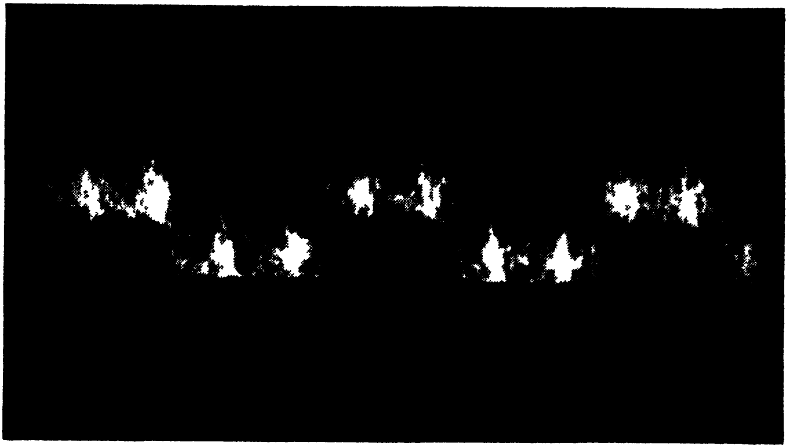
## 58. Signal to noise ratio and optimum band-width

The upper limit to the *sensitivity* of a receiver is determined by the amount of electrical noise produced in its input circuits and first valve stages; wanted signals that are received at such a low level as to be comparable to these minute noise voltages represent the smallest detectable signals, and nothing is to be gained by the addition of further amplifier stages, since these will amplify both signal and noise together. These noise voltages arise in the aerial and in any resistance element in the circuit, due to random movement of electric charges by thermal agitation, and also they arise in the valves themselves due to the random nature of the electron emission from their cathodes.[20, 21] In practice it is usually the input circuit and first valve which produce the significant amount of noise since these have the greatest amplification following them. Sometimes, however, especially with very high-frequency superheterodyne receivers, the first stage has a low gain and an appreciable amount of noise may arise from the second stage, together with a contribution from the local oscillator valve. The total amount of noise produced depends, to some extent, on the design of the early stages and coupling circuits.[19, 24]

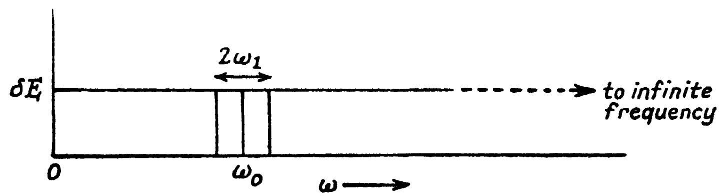
The waveform of a noise voltage is said to be completely random, by which is meant that, if measured over any period of time it never repeats itself. If viewed on a cathode-ray oscilloscope it presents a spiky, fluttering appearance, as shown in the photograph of the noise at the output of a receiver superposed on an injected square-wave signal (Fig. 90 (a)). It is usual to measure the ultimate sensitivity of a receiver by injecting a square-wave (or square-wave modulated carrier) signal from a calibrated signal generator and adjusting its level until it is equal to the observed noise level. This observation [19] may be made either on a cathode-ray oscilloscope as in the figure, in which case it becomes a subjective, though in practice fairly accurate, process, or may be made on a thermal meter which reads true R.M.S. values.

In spite of its random waveform a noise voltage has a definite R.M.S. value and gives a steady reading on a thermal meter. It

may be said to possess a continuous spectrum of *uniform amplitude* but random phase sinusoidal components, in which it differs from the spectrum of an infinitesimally short pulse only in the randomness of the phases (see Fig. 90 (b)). This means that if we could pass the noise energy through an ideal band-pass filter of constant band-width $\pm\omega_1$ but variable mid-band frequency $\omega_0$, the energy at the output of the filter would be constant, independent of $\omega_0$. More-



(a)




(b) Amplitude Spectrum—Phase Spectrum is Random.

Fig. 90.—Electrical Noise—Its Waveform and Spectrum.

over, if the band-width were varied the amount of energy passed would always be proportional to this band-width. The noise produced in any circuit arises in the resistive elements only, and not in the reactive, though of course in high-frequency circuits resistance may appear in nominally reactive components due to various losses. The R.M.S. value of the noise voltage $E_n$ appearing across the terminals of a resistance $R$ ohms depends on the value of $R$ and also upon the *absolute* temperature of the element,

$T^0K$; the exact expression has been calculated by Johnson and Nyquist [20]:

$$E_n = \sqrt{(4kTBR)} \text{ volts, R.M.S.} \quad . \quad . \quad . \quad (268)$$

where $B$ is the band-width in cycles/sec. over which the measurement is made and $k$ is Boltzmann's constant, equal to $1{\cdot}380 \times 10^{-23}$ joules/degree.

The noise voltage produced between the anode and cathode of a valve is of the same general character, so that a valve may be said to possess an equivalent resistance, equal to that resistance value which would give the same noise voltage. It is usual to "refer this resistance to the valve grid," so that the valve amplification is included. In this way the effective noise resistance of the valve is considered as a resistance element between the valve grid and cathode, at the same absolute temperature as the rest of the grid input circuit.

The band-width $B$ in the above equation assumes an idealised flat-topped filter—the phase characteristic is immaterial—and we have already stated that it does not matter what the mid-band frequency is, the noise is proportional to $\sqrt{B}$ so long as a uniform noise spectrum may be assumed. However, practical filter characteristics are not flat-topped and for calculation of their effect on the noise voltage it is convenient to take a very narrow value for $B$ so that we may write $B \to \delta\omega$ in the above equation.

$$\delta E_n = \sqrt{(4kTR\delta\omega)} \quad . \quad . \quad . \quad . \quad . \quad (269)$$

Now the spectrum of any signal that it may be required to receive, and which together with the noise voltage passes through the selective circuits of a receiver, is not uniform (unless the signal be an infinitesimal pulse) and so it may be expected that varying the band-width of these selective circuits will have different effects on the signal and on the noise voltage. In fact we shall see that there may be a critical band-width, giving a maximum signal/noise ratio, if this ratio be measured in a certain way.

If $|Z(\omega)|$ is the modulus of the transfer characteristic of a receiver, measured as the ratio of the voltage across the output terminals to the voltage across the input terminals, and if it be assumed that all the noise appears in a resistance $R$ across the input terminals and none arises in later stages of the receiver, then the output noise is easily calculated. This characteristic may be considered to be split up into adjacent narrow bands of width $\delta\omega$, each of which passes a component of noise given be 269; the total noise *power* output of

the receiver is then proportional to the sum of all these components from zero to infinite frequency:

$$\text{Noise } power \text{ output of receiver} \propto 4kTR\int_0^\infty |Z(\omega)|^2 \, . \, d\omega \quad . \quad . \quad (270)$$

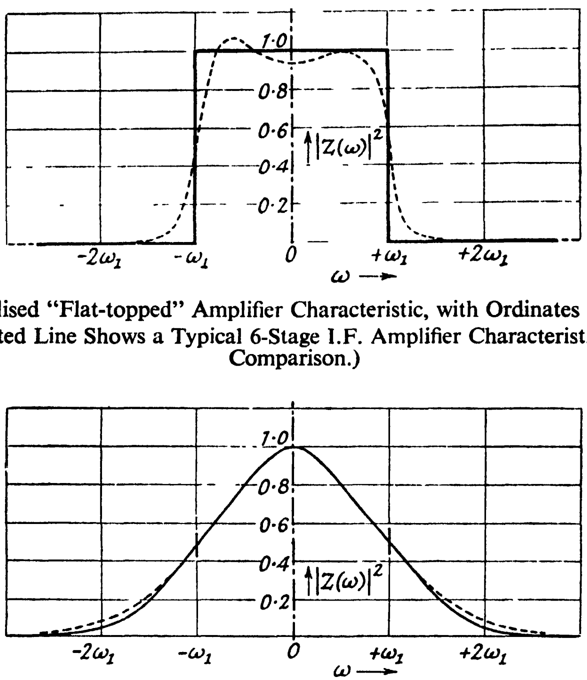
that is, proportional to the area of the receiver characteristic modulus squared.[24] The phase-shift characteristic has no effect on this noise owing to the random nature of its phase spectrum.

Then if two receivers, having different forms of transfer characteristic (which may include absolute levels of gain), have identical input conditions, that is to say identical resistances coupled to the input terminals in identical ways, then their respective output noise powers will be in the ratio of the areas of their characteristic moduli squared. It is interesting to consider two idealised cases which closely approach practical conditions—the flat-topped and the "probability" function characteristics, shown respectively in Fig. 69 and Fig. 75 (a), curve (2). These curves have already been treated (Sec. 53) as limiting characteristics of an amplifier or receiver containing a very large number of stages, and most practical characteristics lie in between these two extreme forms (see Sec. 54). The latter is a very close approximation to a multi-stage tuned-circuit amplifier characteristic as has been illustrated by Fig. 82 and discussed in Sec. 53.

These two idealised characteristics have been plotted in Fig. 91 with their ordinates squared; the "band-width" of the probability curve has been made equal to that of the flat-topped, $\pm\omega_1$, at the half power points. Superposed on the idealised flat-topped characteristic there is shown, in this figure, the characteristic of a six-stage amplifier which the author has had occasion to use, using coupled tuned circuit loads of the type shown in Fig. 87 (b); this provides an illustration of the method, discussed in Sec. 46, of choosing the band-width of an idealised characteristic which is intended to represent a practical one (the areas of the $|Z(\omega)|$ diagrams have been made equal as well as their average levels over the tops of the characteristics). A measurement of the areas of the $|Z(\omega)|^2$ diagrams in Fig. 91 (a) shows a difference of only 1 per cent., so that for the purpose of noise calculation the idealised characteristic is very accurate. Also, superposed on the curve in Fig. 91 (b) (the "probability" curve with ordinates squared) is the characteristic of a six-stage receiver or amplifier with tuned-circuit loads, also with its ordinates squared, and these curves are again

extremely close to one another, the areas differing by 3 per cent.

Measurement of the enclosed areas of these idealised curves (with a planimeter, or otherwise) shows that if the flat-topped characteristic be taken as $(\omega_1 \times 100)$ units, the "probability" curve characteristic is $(\omega_1 \times 103{\cdot}5)$ units. Their respective noise power outputs therefore differ by only 3·5 per cent., and it is very likely that any




(a) Idealised "Flat-topped" Amplifier Characteristic, with Ordinates Squared. (Dotted Line Shows a Typical 6-Stage I.F. Amplifier Characteristic for Comparison.)



(b) Idealised "Probability" Curve Characteristic with Ordinates Squared. (Dotted Line Shows Characteristic of 6-Stage Tuned-circuit Amplifier for Comparison.)

Fig. 91.—Noise Output of Two Idealised Amplifiers. Curves of $|Z(\omega)|^2$.

practical amplifier with the same nominal band-width at the half-power points would have a characteristic lying somewhere in between these two extreme idealised forms and its noise output would also be very much the same; thus, variation of characteristic shape does not much change the noise power output. Obviously characteristics may be conceived which lie outside these idealised forms, for example a badly adjusted I.F. amplifier, using coupled

tuned-circuit loads, may be arranged to have "peaks" on its characteristic similar to those in the dotted curve in Fig. 91 (a) but much more pronounced. Such a characteristic may give a slightly different noise output, but is not of much practical use and need not be included here.

The response of an amplifier to any wanted signal having a definite amplitude and phase spectrum, as opposed to the random spectrum of a noise voltage, depends on the band-width and shape of the transfer characteristic. It is interesting to see how the response to a signal may be distinguished from the noise output and how the discrimination depends on the band-width, since this is often of great importance in radar and television (wide band) receiving amplifiers. For this purpose it is convenient to take as the wanted signal a rectangular pulse (or picture element) of duration $T_1$ (see Fig. 77 (a)) and to estimate how the signal-to-noise ratio varies at the output of an amplifier as its band-width is changed in relation to $1/T_1$, using for the amplifier transfer characteristics these two idealised forms of Fig. 91—the flat-topped and the "probability" curve. The phase-shift characteristic of the amplifier would have, in general, an influence on the response to such a signal, but we may regard the idealised phase-shift characteristics as linear here. This is justified if we assume these idealised characteristics to be those of an amplifier containing a large number of stages, since in order to produce such a receiver, having a band-width $\omega_1$ at any arbitrary level (say at half-power points) each stage alone must have a band-width much greater than $\omega_1$. Consequently the final pass-band of the whole amplifier covers a much more linear portion of the phase-shift characteristic than does that of a single stage amplifier. This point has already been discussed in Sec. 53 and illustrated by Fig. 82, where it has been shown that six stages may be taken as a "large number" for practical purposes.

The responses of both these idealised characteristics to an applied step wave have been calculated already; thus Fig. 70 shows the response of the flat-topped transfer characteristic, being the sine-integral function given by equation 270, while Fig. 75 (b), curve (2), shows the response of the "probability" curve characteristic, which is itself of the form of the "probability-integral" discussed in Sec. 53 and given by equation 237.

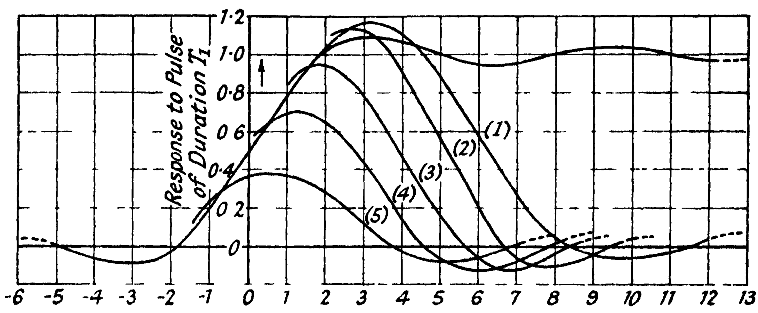
Both these response curves have been plotted again in Fig. 92, together with the responses to rectangular pulse waves of various durations $T_1$, which may be related to the reciprocal of the

band-width ($2\pi/\omega_1$ seconds). These responses have been obtained from the step-wave responses, $h(t)$, by superposition, as explained in Sec. 42:

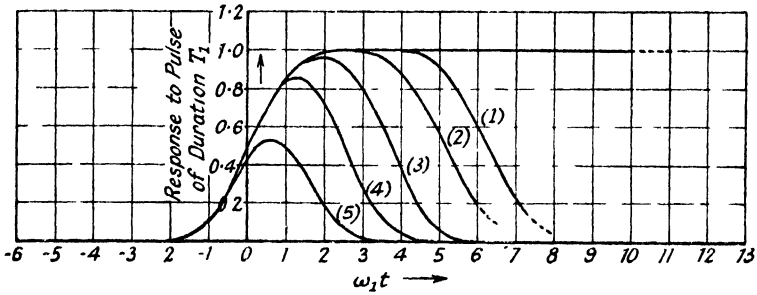
Response to rectangular pulse of duration $T_1$

$$=h(t)-h(t-T_1) \quad . \quad . \quad . \quad . \quad . \quad [(214)]$$

Thus these responses, as drawn in the figures, may be interpreted




(a) Response of Flat-topped Characteristic.




(b) Response of "Probability" Curve Characteristic.

Fig. 92.—Responses of Flat-topped and "Probability" Curve Characteristics to Rectangular Pulses of Duration $T_1$.

[Where $T_1$=(1) 1·0× (2) 0·8× (3) 0·6× (4) 0·4× (5) 0·2×$2\pi/\omega_1$.]

for any practical pulse-width and band-width, within the chosen limitations. There are many ways of assessing the magnitude of a signal, but for our present purpose of determining the band-width which gives the maximum signal-to-noise ratio, with applications to television and radar receivers in mind, it is preferable to use the *peak* of the response signal and the R.M.S. value of the noise.

The peak values of the response signals corresponding to various band-widths may be read off the curves in Fig. 92, and we have already estimated the noise power outputs by measuring the areas under the curves in Fig. 91; then the R.M.S. noise voltages are proportional to $\sqrt{(\omega_1 \times 100)}$ for the flat-topped characteristic and to $\sqrt{(\omega_1 \times 103{\cdot}5)}$ for the "probability" curve characteristic.

Curves of peak signal/R.M.S. noise voltage ratio have been plotted in Fig. 93, for both types of transfer characteristic, against the half band-width $\omega_1$. From these it may be seen that there is an optimum band-width of $\omega_1 = 4{\cdot}5/T_1$ for the flat-topped and $2{\cdot}6/T_1$

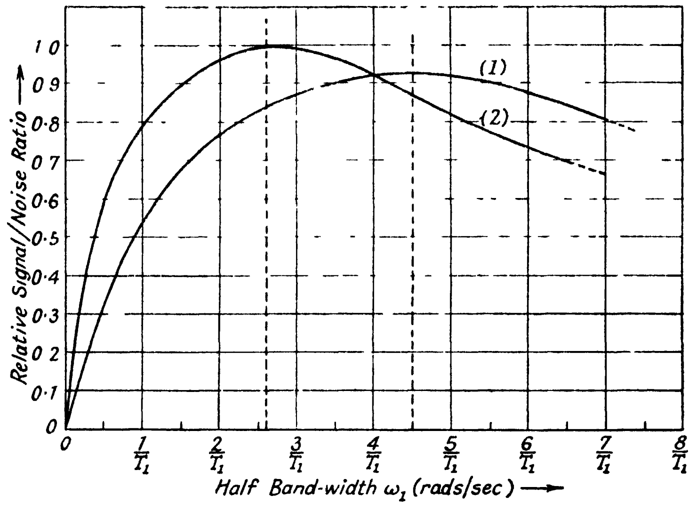



Fig. 93.—Relative Peak Signal/Noise Ratios with (1) Flat-topped and (2) "Probability" Curve Characteristics.

for the "probability" curve characteristic. With these band-widths the received signal, corresponding to an applied pulse of duration $T_1$, will stand out above the noise by the maximum amount.

These curves must not be interpreted with too great an accuracy. They have been estimated, it must be remembered, from idealised responses, and it is likely that most practical receivers will require a band-width lying somewhere between the two optimum values. Neither maxima are very critical in these curves, and band-widths may be varied by 100 per cent. before the signal/noise ratio falls by 10 per cent.

## REFERENCES

1. O. S. PUCKLE, "The Transient Aspect of Wideband Amplifiers," *Wireless Engineer*, Vol. 12, p. 251, May 1945.
2. G. W. EDWARDS and E. C. CHERRY, "Amplifier Characteristics at Low Frequencies, etc.," *J.I.E.E.*, Vol. 87, No. 524, Aug. 1940.
3. H. A. WHEELER, "Wideband Amplifiers for Television," *Proc. I.R.E.*, p. 429, July 1939.
4. G. BUILDER, "The Amplification of Transients," *Wireless Engineer*, Vol. 12, p. 246, May 1935.
5. C. W. OATLEY, "The Distortionless Amplification of Electrical Transients," *Wireless Engineer*, Vol. 8, p. 244, May 1931.
6. S. BUTTERWORTH, "On the Theory of Filter Amplifiers," *Wireless Engineer*, Vol. 7, p. 536, Oct. 1930.
7. H. E. KALLMAN, R. E. SPENCER, and C. P. SINGER, "Transient Response," *Proc. I.R.E.*, Vol. 33, No. 3, p. 169, March 1945.
8. E. K. SANDEMAN, "Coupling Circuits as Band-pass Filters," *Wireless Engineer*, Pt. 1, Sept. 1941, p. 361; Pt. 2, Oct. 1941, p. 406; Pt. 3, Nov. 1941, p. 450; Pt. 4, Dec. 1941, p. 492.
9. C. B. AIKEN, "Two-Mesh Tuned Coupled Circuit Filters," *Proc. I.R.E.*, Vol. 25, No. 2, p. 230, Feb. 1937.
10. F. E. TERMAN, "Radio Engineering," Chap. 3, p. 72, McGraw-Hill, New York, 1937.
11. D. WEIGHTON, "Performance of Coupled and Staggered Circuits in Wide Band Amplifiers," *Wireless Engineer*, Vol. 21, p. 468, Oct. 1944.
12. E. K. SANDEMAN, "Generalised Characteristics of Linear Networks," *Wireless Engineer*, Vol. 13, p. 637, Dec. 1936.
13. A. V. HAEFF, "A Wideband Inductive Output Amplifier," *Proc. I.R.E.*, Vol. 28, p. 126, March 1940.
14. V. K. ZWORYKIN, G. A. MORTON, and L. MALTER, "The Secondary Emission Multiplier—a New Electronic Device," *Proc. I.R.E.*, Vol. 24, No. 3, p. 351, March 1936.
15. A. V. BEDFORD and G. L. FREDENDALL, "Transient Response of Multi-stage Video-frequency Amplifiers," *Proc. I.R.E.*, Vol. 27, p. 277, April 1939.
16. N. W. MCLACHLAN, "Reproduction of Transients by Television Amplifiers," *Wireless Engineer*, Vol. 13, p. 519, Oct. 1936.
17. J. CARSON and O. ZOBEL, "Transient Oscillations in Electric Wave Filters," *Bell Syst. Tech. Jour.*, Vol. 2, p. 1, July 1923.
18. J. CARSON, "Electric Circuit Theory and the Operational Calculus," Part II, *Bell Syst. Tech. Jour.*, Vol. 5, p. 50, Jan. 1926.
19. D. O. NORTH, "The Absolute Sensitivity of Radio Receivers," *R.C.A. Review*, Vol. 6, p. 332, 1941–2.
20. H. NYQUIST, "Thermal Agitation of Electric Charge in Conductors," *Physical Review*, Vol. 32, p. 110, July 1928.
21. D. A. BELL, "Fluctuations of Electric Current," *J.I.E.E.*, Vol. 93, Part III, No. 21, p. 37, Jan. 1946 (see included list of references).
22. W. R. MACLEAN, "Ultimate Bandwidths in High-Gain Multi-stage Video Amplifiers," *Proc. I.R.E.*, Vol. 32, p. 12, Jan. 1944.

23. J. E. Cope, "Some Considerations in the Design of Wide-Band Radio-Frequency Amplifiers," *J.I.E.E.*, Vol. 92, Part III, No. 20, p. 237, Dec. 1945.
24. F. B. Llewellyn, "A Study of Noise in Vacuum Tubes and Attached Circuits," *Proc. I.R.E.*, Vol. 18, No. 2, p. 243, Feb. 1930.
25. H. S. Black, "Stabilised Feedback Amplifiers," *Electrical Eng.*, Vol. 53 p. 114, Jan. 1934.

The following are also of interest:

(A) C. C. Eaglefield, "Square Wave Analyser," *Wireless Engineer*, Vol. 22, May 1945 (build-up time and overswing measurement).
(B) C. E. Lane, "Phase Distortion in Telephone Apparatus," *Bell Syst. Tech. Jour.*, Vol. 9, p. 493, July 1930 (phase distortion—see also the bibliography).
(C) H. E. Kallman, "Portable Equipment for Observing Transient Response of Television Apparatus," *Proc. I.R.E.*, Vol. 28, Aug. 1940.
(D) V. D. Landon, "Cascade Amplifiers with Maximal Flatness," *R.C.A. Review*, Vol. 5, p. 347, 1941.
(E) N. R. Campbell and V. J. Francis, "A Theory of Valve and Circuit Noise," *J.I.E.E.*, Vol. 93, Part III, No. 21, p. 45, Jan. 1946.

# CHAPTER 7

# ASYMMETRIC SIDEBAND CHANNELS

## 59. Single-sideband working—its applications and general effects

It has been shown in Chapter 2, Sec. 15, that if a carrier wave is modulated in amplitude by a signal the resulting wave consists of a spectrum of sinusoidal components (the sidebands) distributed symmetrically about a central term of carrier frequency, and furthermore that the spectrum is the same as that of the modulating signal itself, except that this spectrum is symmetrical about *zero* frequency. The band-width of the modulated carrier spectrum depends on the modulating signal, increasing as the frequencies contained in this signal rise; for example, speech channels may require band-widths of only a few Kc/s, whereas television and pulse channels, which are required to transmit extremely sharp transients and impulses, require band-widths of several Mc/s. There are advantages to be obtained by reducing the band-width used in certain types of communication system, especially the very wide-band systems, and this is sometimes done by filtering out all, or part, of the sideband components lying above (or below) the carrier frequency; such methods of communication are known as single-sideband or asymmetric sideband systems.

Since the spectrum of an amplitude-modulated wave is symmetrical about the carrier frequency, it is fairly true to say that the information contained in the upper and in the lower sidebands is the same, but the signal that can actually be extracted from one or other of these sidebands, when its partner is filtered out, may be distorted. The problem of this distortion has received the attention of many writers concerned both with television [1, 2, 3, 4, 7, 8, 10] and speech channels,[5, 6] and also with line telegraphy problems.[9] In this chapter we shall investigate the distortion of a modulated wave, due to the total or partial suppression of one sideband, in such a way that the results may be applicable to any communication system. The effects of sideband asymmetry require study for three purposes:

(1) To find the distorting effects of tuning the carrier frequency to one edge of a filter, so that one sideband is practically suppressed. (Single-sideband working.)

(2) To find the effects of mistuning the carrier from the mid-band frequency so that the sidebands are made asymmetric in amplitude and phase, but neither is completely suppressed. (Asymmetric-sideband working.)

(3) To find the effects of the inherent asymmetry of all band-pass structures (see Chapter 3, Secs. 30, 31, and 33) which, though rather of academic interest, is of importance in connection with the band-pass/low-pass analogy and the transfer of a carrier frequency to zero.

It will be seen that the general effects of disturbing the symmetrical arrangement of the sidebands of an amplitude-modulated wave are: (1) to distort the envelope (and hence the wanted signal) waveform; (2) to introduce phase modulation into the carrier. These distortions depend in detail on the waveform of the modulating signal, and in general the degree of distortion *increases with the depth of modulation*, so that a non-linear relation enters into the problem. For this reason the distortions cannot be overcome by the use of correcting networks, although it is possible to minimise the undesirable effects by taking certain precautions. Before starting on an analysis, let us look at various types of communication systems to see the possible advantages or disadvantages of sideband suppression.

Single-sideband working has been used for carrier line telephony for many years, the chief advantages in this field being that double the number of channels are made available in a given range of frequency and greatly increased amplifier gain or economy is possible. In this case little or no distortion need arise due to sideband suppression for reasons which depend mainly on the fact that modulation frequencies do not extend down to zero frequency, but only to a certain fixed limit. Thus a clear frequency range is left between the carrier and the innermost sideband components (see Fig. 94 (a)) and filters may be designed with suitably steep cut-off characteristics which substantially eliminate one sideband. Means may then be provided for increasing the amplitude of the received carrier component, thereby effectively reducing the depth of modulation—which, as we shall see later, has the effect of reducing the distortion of the envelope and hence of the received signal.

In the case of broadcast (sound) channels the problem of single-sideband working is fundamentally different, since much higher carrier frequencies are used than for carrier line telephony and, at

the same time, lower modulation frequencies are required, probably extending down to 30 c/s (see Fig. 94 (b)). This means that the complete suppression of one sideband becomes less easy owing to the difficulty of designing filters with suitably sharp rates of cut-off, and attention has been given by a number of investigators [5, 6] to the problem of suppressing one sideband only partially (asymmetric- or vestigial-sideband working). However, most published work on this problem deals only with pure tone modulation, of a single frequency, ignoring the essential transient nature of speech and music signals.

If asymmetric sideband systems are found satisfactory in practice, for speech and music, it is for two reasons. First, most of the modulation energy lies in the lower frequencies, below about 2 Kc/s, so that a great number of the corresponding sidebands lie in the

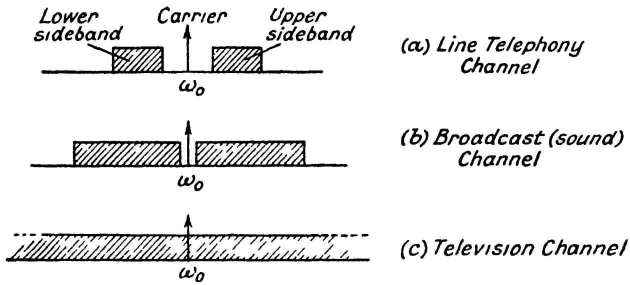



Fig. 94.—Sideband to Carrier Frequency Spacings.

region bordering on the carrier; parts of both the upper and the lower sidebands in this region are transmitted, which minimises the distortion. Secondly, the ear is not sensitive to phase distortion, so that it is not important to preserve the exact *shape* of the carrier envelope, but only the relative amplitudes of the individual components it contains.

In the case of television broadcasting, modulation frequencies down to zero frequency are used, so that the resulting sidebands approach and touch the carrier component (see Fig. 94 (c)). Television signals are essentially of a transient nature and it is necessary that the envelope transient shapes should be preserved as well as possible, since the eye, unlike the ear, is sensitive to changes of wave shape whether produced by phase-shift or by amplitude distortion. There are possible advantages to be gained by the use of asymmetric sideband systems for television and also for pulse

communication, which presents a similar problem: (a) economy in channel band-width required; (b) if this advantage be sacrificed, as an alternative, higher modulation frequencies may be accommodated in existing band-widths, resulting in improved picture definition; (c) increased stage-gain in the receiver amplifiers is possible, due to the reduction in band-width (see Sec. 56); (d) an increased signal/noise ratio results if both the transmitter and receiver are worked on the asymmetric sideband system, since all the transmitter power can be put into half the band-width, resulting in increased received signal strength, while the noise power is halved in the receiver (see Sec. 58); (e) the effects of external interference in the receiver are reduced by the reduced receiver band-width.

Since there is usually no frequency gap in such channels between the carrier component and the sidebands the use of practical band-pass filters would imply that, at the best, only partial suppression

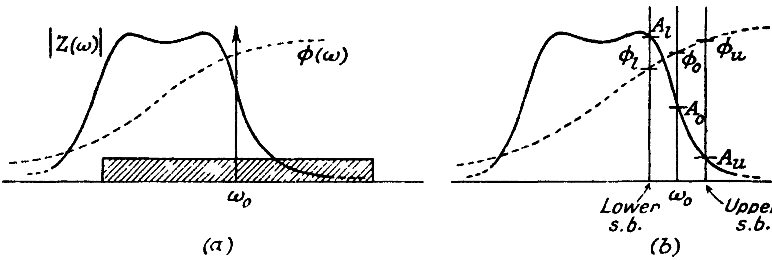



Fig. 95.—Asymmetric Sideband Channels.
(a) Television Signal. (b) Simple Sinusoidal Signal.

of the innermost sidebands can be achieved, corresponding to the lower modulation frequencies.

The cut-off characteristic of the band-pass filter, or filters, through which the modulated wave is passed will distort these innermost sidebands in amplitude and phase and this results in distortion of the detected signal envelope. Fig. 95 (a) illustrates a signal tuned to one edge of a band-pass characteristic, in this case on the high-frequency side, though the low-frequency side could alternatively be used; one sideband is transmitted and the other partially attenuated, the sidebands close to the carrier merely being distorted asymmetrically. Since it is the sidebands close to the carrier that are distorted in this way, it will show up as low-frequency distortion of the carrier envelope, an effect which can possibly have serious consequences in television reception and which must be watched

closely in any analysis made. It is not likely to be of such importance for pure pulse communication channels. At the same time the virtual removal of the outer sidebands from those on one side of the carrier must affect the high-frequency components of the carrier envelope, which are important in both television and pulse channels, or in any other channel in which preservation of build-up time is essential.

If the carrier frequency, in Fig. 95 (a), had been tuned to the mid-band frequency of the band-pass characteristics, it would have been possible to have transferred this frequency to zero in order to determine directly the resulting envelope distortion (see Sec. 41), assuming these characteristics to be perfectly symmetrical about mid-band. In that case there would have been no carrier phase distortion and the envelope distortion would have been the same as that suffered by the envelope wave (i.e. the signal) in passing through a low-pass filter having the "equivalent envelope-frequency characteristics." Fig. 64 has already been used to illustrate this equivalence. However, in the present case in which the carrier frequency is tuned to one side of the channel band-pass characteristics, this theorem cannot be applied as it stands and some extention to the principle must be sought. In the following pages this line of thought is followed out and means are developed for finding the "equivalent envelope-frequency characteristics" which are applicable to any form of modulating signal and any depth of modulation, and supported by illustrative examples.

## 60. Modulation by a pure tone

Modulation by a purely sinusoidal wave is the simplest type, involving only two sidebands, and it provides a useful means of observing the mechanism by which envelope and carrier distortion arise when the symmetry of these sidebands is disturbed. Such modulation has been given prominence in published analyses and the results can be rather misleading, since, for one thing, some of these publications derive results which are applicable to small depths of modulation only; furthermore, it is unwise to base general conclusions on the use of sinusoidal modulation, since this does not, strictly speaking, constitute a signal, because it cannot convey information, which essentially must involve a *change* of some kind. Let us therefore examine the asymmetric sideband distortion of such a wave, with this limited aspect in mind.

If a carrier wave $E \cos \omega_0 t$ be modulated in amplitude by a

sinusoidal envelope $(mE)\cos\omega t$, the resulting wave is given by the expression:

$$e_1 = E[1+m\cos\omega t]\cos\omega_0 t \quad . \quad . \quad . \quad (271)$$

as has already been shown in Sec. 15. In using the above expression we have assumed, without loss of generality, that the time origin corresponds to the peak of the carrier and also the peak of the envelope. This eliminates any initial phase angles, which serve no purpose in the analysis and which merely lengthen the expressions involved. The expression (271) above separates into carrier, upper, and lower sideband terms:

$$e_1 = E\cos\omega_0 t + \frac{mE}{2}\cos(\omega_0+\omega)t + \frac{mE}{2}\cos(\omega_0-\omega)t \; . \quad (272)$$

Let this wave be applied to the input terminals of the asymmetric sideband channel whose transfer impedance characteristics are shown in Fig. 95 (b). The carrier and upper and lower sideband voltages, across the output terminals, are altered in amplitude by the factors $A_0$ $A_u$ and $A_l$ and in phase by $\phi_0$, $\phi_u$ and $\phi_l$ respectively,* so that the wave on the output terminals is represented by:

$$e = E \,.\, A_0 \,.\, \cos(\omega_0 t+\phi_0) + \frac{mE}{2} \,.\, A_u \,.\, \cos[(\omega_0+\omega)t+\phi_u]$$

$$+\frac{mE}{2} \,.\, A_l \,.\, \cos[(\omega_0-\omega)t+\phi_l] \; . \quad (273)$$

The factors $A_l$, $A_u$, $\phi_l$ and $\phi_u$ are dependent on the frequency $\omega$.

This modulated wave may be represented by the sum of three vectors, both before and after its passage through the channel. We have already shown (Fig. 20) the vector diagram for an undistorted wave, consisting of three vectors with angular velocities $\omega_0$, $(\omega_0+\omega)$, and $(\omega_0-\omega)$ radians/sec. respectively, which represent the carrier, the upper, and the lower sideband terms. The carrier vector may be regarded as being stationary on the paper, while the sideband vectors rotate in opposite directions with angular velocities of $\pm\omega$ radians/sec. This vector diagram has been drawn again in Fig. 96 (a), together with the diagram (b) representing the wave after its transmission through the asymmetric sideband channel. This shows, in accordance with equation 273, the three vectors modified

* These factors are, of course, functions of frequency, and should be written $A_u(\omega)$, $\phi_u(\omega)$ . . . , etc., but the $(\omega)$ is omitted to simplify the equations that follow, and it should be understood here.

in length and phase-angle. The instantaneous amplitude of the modulated wave is given by the sum of these three vectors; in the case of the undistorted wave, (a), this resultant will lie along the line of the carrier vector, since the sideband vectors are always symmetrically disposed about it; but in the case of the distorted output wave, (b), the resultant vector will not coincide with the carrier vector.

Thus the resultant vector may be divided into two orthogonal components, one in phase with the carrier and one in quadrature

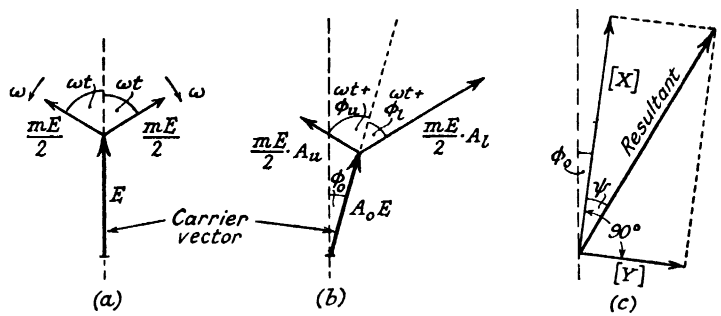



Fig. 96.—Distortion of Sideband Components due Asymmetry of Channel Characteristic.

with it. These components may be found by writing equation 273 as:

$$e=\left[A_0+\frac{mA_l}{2}\cos(\omega t+\phi_0-\phi_l)+\frac{mA_u}{2}\cos(\omega t+\phi_u-\phi_0)\right].E\cos(\omega_0 t+\phi_0)$$
$$+\left[\frac{mA_l}{2}\sin(\omega t+\phi_0-\phi_l)-\frac{mA_u}{2}\sin(\omega t+\phi_u-\phi_0)\right].E\sin(\omega_0 t+\phi_0) \quad . \quad (274)$$

The terms in square brackets represent the magnitudes of the envelopes of the *in-phase* and *quadrature* components of the distorted wave. Denoting these by [$X$] and [$Y$] respectively, the wave (274) is given by:

$$e=[X].E\cos(\omega_0 t+\phi_0)+[Y].E\sin(\omega_0 t+\phi_0) \quad . \quad (275)$$

These two components are shown in Fig. 96 (c). The resultant

distorted wave is given by adding these two orthogonal components together:

$$e=E\,.\,\sqrt{(X^2+Y^2)}\,.\cos\left(\omega_0 t+\phi_0+\tan^{-1}\frac{Y}{X}\right)$$

$$=e_E\,.\cos\,(\omega_0 t+\phi_0+\psi)\quad .\ .\ .\ .\ .\ .\ .\ .\quad (276)$$

Thus $e_E$, the envelope of this wave which has been distorted by sideband asymmetry, is given by the expression:

$$e_E=E\,.\,\sqrt{(X^2+Y^2)}\quad .\ .\ .\ .\ .\quad (277)$$

and the carrier phase modulation $\psi$ by:

$$\psi=\tan^{-1}\frac{Y}{X}\quad .\ .\ .\ .\ .\ .\quad (278)$$

As the carrier and sideband vectors, in Fig. 96 (b), rotate with time at their respective angular frequencies, their projections [$X$] and [$Y$], along the carrier axis and orthogonal to it, vary in length. The resultant of these components [$X$] and [$Y$] represents the distorted output wave in magnitude and phase. It is interesting to see how the magnitude and phase vary in a particular case, which is most easily shown by plotting the locus of the resultant vector for a complete envelope cycle. Suppose then, for example, that the characteristics $A_l$, $\phi_l$, $A_u$, . . . etc. (Fig. 95 (b)) have the following numerical values:

$$\left.\begin{array}{lll}\text{Carrier component:} & A_0=0{\cdot}6, & \phi_0=120^\circ\\ \text{Upper sideband:} & A_u=0{\cdot}25, & \phi_u=170^\circ\\ \text{Lower sideband:} & A_l=1{\cdot}0, & \phi_l=20^\circ\end{array}\right\}\quad .\ .\quad (279)$$

and that the depth of modulation is 100 per cent., i.e. $m=1{\cdot}0$. Substitution in equation 274 gives the values of [$X$] and [$Y$]. Fig. 97 shows a plot of the resultant vector locus, which in this case of pure sinusoidal modulation must be an ellipse. The ringed points on this locus correspond to positions of the resultant vector at equal time increments of $\omega t=20^\circ$.

From this locus the envelope amplitude $e_E$ (resultant vector length) and carrier phase angle $\psi$ may be read off. Fig. 98 (a) gives a plot of the distorted wave envelope and (b) of the carrier phase modulation, over one envelope cycle. In normal communication channels it is the wave envelope which is of importance, and from this diagram the degree of harmonic distortion introduced by the sideband asymmetry in the example may be judged.

The forms of the envelope and carrier phase modulation are readily found, in the general case, by substituting for [X] and [Y] in 277 and 278 from equation 274, giving:

The envelope, $e_E = E\Big[A_0^2 + \frac{m^2}{4}\{A_u^2 + A_l^2 + 2A_uA_l \cos(2\omega t + \phi_u - \phi_l)\}$

$$+ mA_0\{A_u \cos(\omega t + \phi_u - \phi_0) + A_l \cos(\omega t + \phi_0 - \phi_l)\}\Big]^{\frac{1}{2}} \quad . \quad . \quad (280)$$

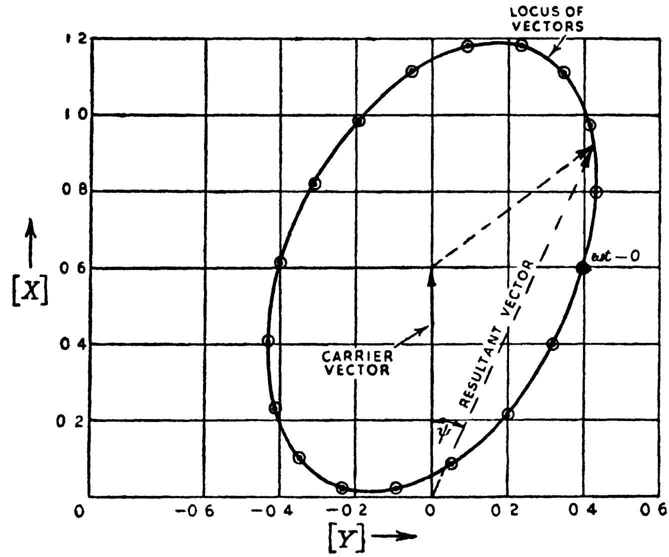


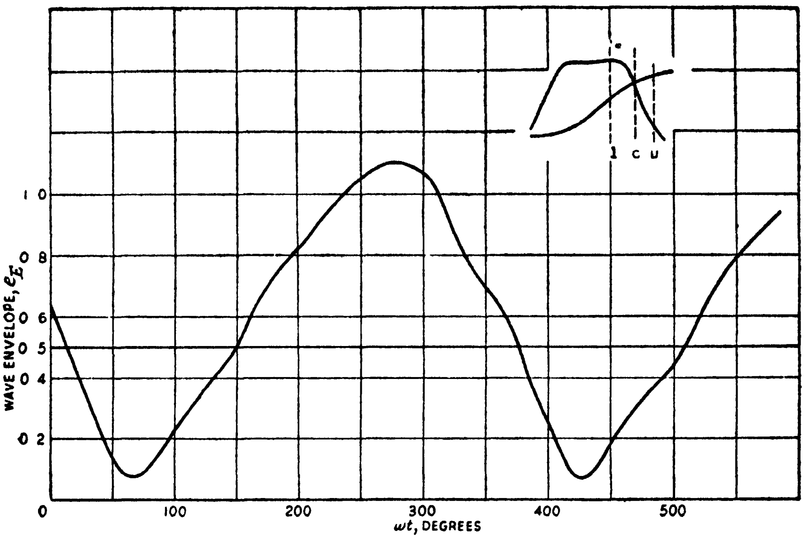

Fig. 97.—Sinusoidal Modulation—Vector Locus.

Carrier phase modulation:

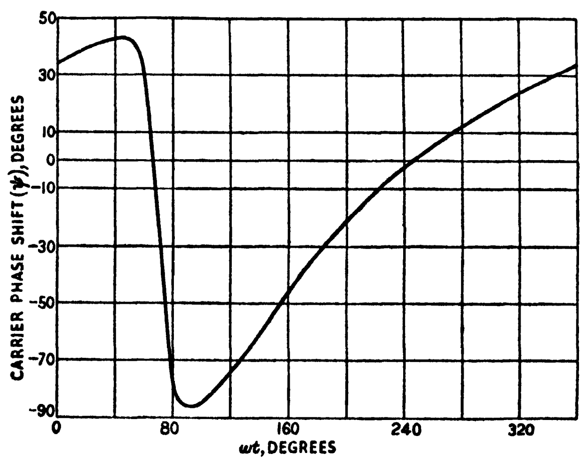
$$\tan\psi = \frac{\frac{mA_l}{2}\sin(\omega t + \phi_0 - \phi_l) - \frac{mA_u}{2}\sin(\omega t + \phi_u - \phi_0)}{A_0 + \frac{mA_l}{2}\cos(\omega t + \phi_0 - \phi_l) + \frac{mA_u}{2}\cos(\omega t + \phi_u - \phi_0)} \quad (281)$$

The above equation (280) for the envelope has been used by Poch and Epstein,[1] who have neglected, however, the terms in $m^2$, so that their analysis applies only to low percentage modulation. But this is the same as neglecting the envelope harmonic distortion altogether, and merely indicates, as the result of sideband asymmetry, a modification to the percentage modulation. Other writers [8] have taken merely the "in-phase" component, [X], of the distorted wave as the resultant output signal, neglecting the

orthogonal component [Y]. This again corresponds to ignoring the envelope harmonic distortion.




(a) Distorted Wave Envelope.




(b) Carrier Phase Modulation.

Fig. 98.—Sinusoidal Modulation—Asymmetric Sideband Distortion.

$A_0 = 0{\cdot}6 \qquad \phi_0 = 120°$

$A_u = 0{\cdot}25 \qquad \phi_u = 170°$

$A_l = 1{\cdot}0 \qquad \phi_l = 20°$

## 61. Single-sideband—one sideband completely suppressed

It can be seen from the vector diagrams in Fig. 96 that the [$Y$] component, which gives rise to the envelope harmonic distortion and the carrier phase modulation, increases as the sideband amplitudes $A_u$ and $A_l$ become more unequal. Then the limiting case in which one sideband is completely suppressed (as in single-sideband systems proper) presents the optimum amount of distortion. Let the upper sideband be completely suppressed, so that $A_u=0$. Then the envelope is, from equation 280:

$$e_E=E\left[A_0^2+\frac{m^2A_l^2}{4}+mA_0A_l\cos(\omega t+\phi_0-\phi_l)\right]^{\frac{1}{2}} \quad . \quad . \quad (282)$$

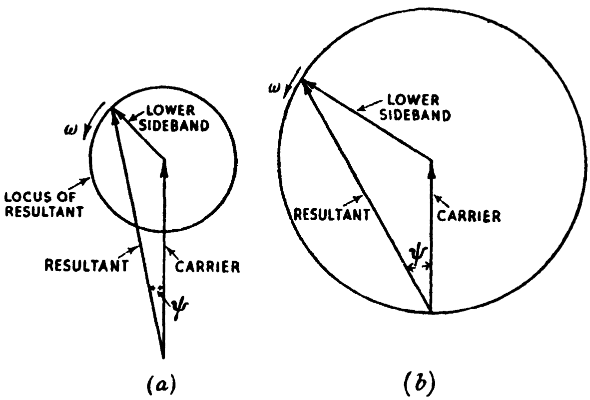



Fig. 99.—Complete Suppression of One Sideband.

The vector diagram for this single-sideband wave is shown in Fig. 99 (a). The length of the resultant vector represents the envelope, $e_E$, while the angle $\psi$ that it makes with the carrier vector gives the carrier phase modulation. The locus of the resultant vector in this case is a circle of radius $mA_l/2$. Fig. 99 (b) shows the vector diagram when the sideband and carrier are equal in magnitude.

Clearly the form of this vector diagram, and hence of the resulting distorted wave, depends on the *ratio* of the two magnitudes involved, $A_0$ and $A_l$. Writing this ratio as $\eta$:

$$\eta=A_l/A_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (283)$$

and inserting in the above equation 282 gives for the envelope:

$$e_E=E\,.\,A_0\left[1+\frac{m^2\eta^2}{4}+m\eta\,.\cos(\omega t+\phi_0-\phi_l)\right]^{\frac{1}{2}} \quad . \quad . \quad (284)$$

This envelope has been plotted, for various values of $m\eta$, in Fig. 100 (a), together with the carrier phase modulation (b) which is given by writing $A_u=0$ in equation 281 and using the ratio $\eta$:

$$\tan\psi=\frac{m\eta\sin(\omega t+\phi_0-\phi_l)}{2+m\eta\cos(\omega t+\phi_0-\phi_l)} \quad . \quad . \quad . \quad (285)$$

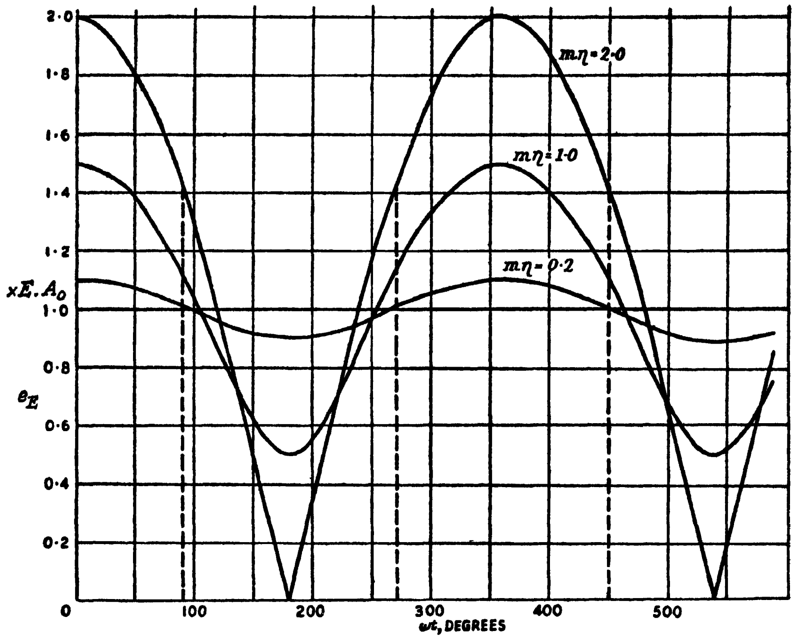
The phase angle $(\phi_0-\phi_l)$ is ignored in this figure. It may be observed how the distortion of what was originally a sinusoidal envelope increases as the depth of modulation $m$ is increased, or alternatively as the unattenuated sideband is increased in amplitude in relation to the carrier component ($\eta$). Distortion is a minimum when the effective depth of modulation of the resulting output wave is kept to a minimum. In the limiting case $m\eta=2$ the sideband and carrier vectors are equal in length, the vector diagram is as shown in Fig. 99 (b), and the resulting envelope just reaches 100 per cent. effective modulation, its waveform being that of a rectified sine wave.

This single sideband modulated wave is, of course, nothing more than the heterodyne between two sine waves of different frequencies, by which description it is probably better known to the reader.[11] The two frequencies are, in this case, those of the carrier and the unattenuated sideband. Thus this wave, which is the resultant of *two* sinusoidal components, is essentially phase modulated, whereas a pure amplitude-modulated wave which has no phase modulation possesses *three* components. The phase modulation, plotted over one envelope cycle in Fig. 100 (b), shows an increase with $m\eta$; it is zero at 0°, 180°, 360° . . . , etc., corresponding to the peaks and the troughs of the envelope, but varies in between these points. The case when $m\eta=2$ is special, as we shall see.

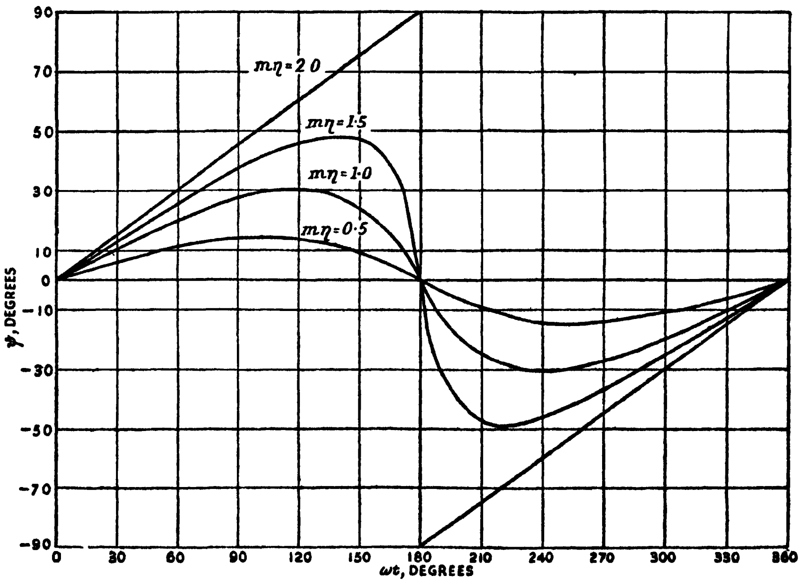
The envelope, given by equation 284, has minimum values (the troughs) when $\cos(\omega t+\phi_0-\phi_l)=-1$, given by:

$$(e_E)_{min}=EA_0\left(1-\frac{m\eta}{2}\right) \quad . \quad . \quad . \quad . \quad (286)$$

at every interval of $\omega t=2\pi$ along the time axis. Now the modulation index $m$ may be anything between 0 and 1·0, while $\eta$ may have any value. Keeping the carrier amplitude $A_0$ constant and varying $A_l$, it can be seen from the vector diagram, Fig. 99, and from the envelope curves, Fig. 100 (a), that while $m\eta<2{\cdot}0$ the envelope trough never reaches zero carrier level; when $m\eta=2{\cdot}0$ we have just 100 per cent modulation of the distorted wave; and when $m\eta>2{\cdot}0$ we have *overmodulation* produced by the single-sideband distortion.




(a) Envelope Waveforms.




(b) Carrier Phase Modulation.

Fig. 100.—Sinusoidal Modulation—Single-Sideband Distortion.

The actual effect of this overmodulation is best illustrated by the vector diagram, Fig. 99. As the sideband vector amplitude $mA_l/2$ is increased until it approaches the carrier amplitude $A_0$ the resultant vector oscillates in phase, about the carrier frequency, by an angle $\psi$ which increases with $A_l$. When $mA_l/2=A_0$ the vectors have the same length and the angle $\psi$ oscillates through $\pm\pi/2$. If $mA_l/2$ is made greater than $A_0$ the vector locus becomes larger in radius than the carrier vector length and an extra cycle is added to the resulting wave for every modulation cycle. That is, the phase modulation

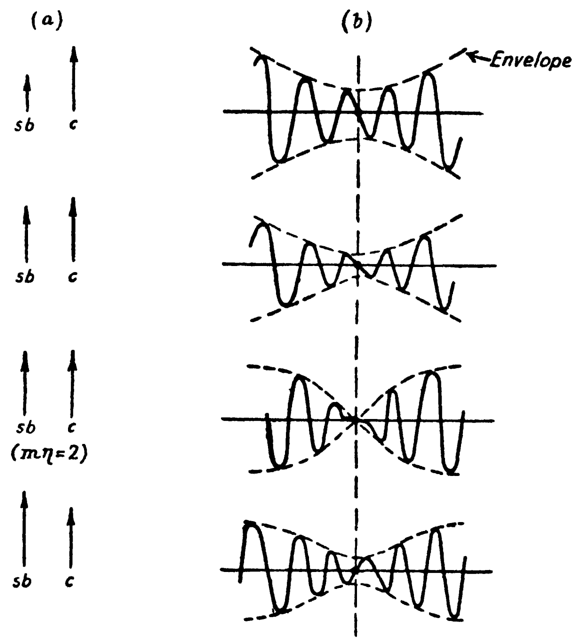



Fig. 101.—Apparent Change in Carrier Frequency.

produces a sudden apparent change in frequency of the resulting modulated wave, as is seen in Fig. 101, which shows the trough of the envelope wave and the manner in which it changes as the remaining sideband approaches, and eventually exceeds, the carrier component in magnitude.

In this particular case of sinusoidal modulation the distortion introduced by the removal of one sideband may be completely corrected by the use of a square-law detector, as may be seen by squaring the expression (284) for the envelope. The apparent depth

of modulation is changed, but the detected signal is theoretically sinusoidal. However, this correction, as we shall see, cannot be applied to other types of modulating waveform and so it is not of much practical interest.

Another method of reducing single- (or asymmetric) sideband distortion, as used in carrier telephony systems, is to reintroduce an amplified carrier component at the receiver end so that the depth of modulation is effectively reduced. In this way the value of $m$ is reduced and the quadrature component $[Y]$ is made smaller than it would otherwise be (see vector diagram, Fig. 96). Thus the carrier phase modulation $\psi$ and the envelope distortion are decreased. The difficulty in this process lies in the correct phasing of the reinserted carrier component, since it must be accurately in phase with the original carrier component at the transmitting end so as to bear the right phase relationship to the sidebands. This is sometimes achieved by filtering out the carrier-frequency component at the transmitting end and sending it via a narrow-band "pilot" channel.

## 62. Equivalent modulation-frequency characteristics

Before continuing with details of the response of asymmetrical sideband channels to transients, let us see whether it is possible to obtain a set of equivalent modulation-frequency characteristics, so that the distortion of the wave *envelope* may be determined in a direct manner from the channel characteristics.

We have already mentioned, in the introductory Sec. 59, that the carrier frequency cannot be transferred to zero in order to obtain the equivalent modulation-frequency characteristics, since this simple theorem can be applied only when the channel characteristics are symmetrical about the carrier frequency (see Sec. 41). Furthermore, it is clear from the last section that the envelope harmonic distortion depends on the depth of modulation, so that the equivalent modulation-frequency characteristics (if they exist) have to be such that the distortion of the response wave varies with the amplitude of the applied envelope wave. In other words, there must be a non-linear element entering into these characteristics.

Actually, a very simple extension may be found to the theorem concerning the transfer of the carrier (mid-band) frequency to zero, whereby it may be applied to asymmetric channel characteristics. We shall see, in the following, that not one but *two* sets of equivalent modulation-frequency characteristics are required, one applying to

the [X] or in-phase component and one to the [Y] or quadrature component. These two distorted components are then determinable by applying the envelope wave to these two sets of characteristics in turn. In the case of a sinusoidal modulation envelope the expressions for the [X] and [Y] components of the response wave are contained in the square brackets in equation 274. It will be noticed that these expressions contain terms referring to the envelope signal (e.g. $m$ and $\omega$) and also terms relating to the channel characteristics (e.g. $A_0$, $A_u$, $\phi_u$ . . . , etc.). These two sets of terms may easily be separated by expanding these expressions for [X] and [Y]. For example, we may write:

$$\cos(\omega t+\phi_0-\phi_l)=\cos\omega t\,.\,\cos(\phi_0-\phi_l)-\sin\omega t\,.\,\sin(\phi_0-\phi_l) \qquad \ldots\ldots\ (287)$$

and so on. The whole expression for the [X] component amplitude then becomes:

$$[X]=A_0\left[\left\{\frac{A_l}{2A_0}\cos(\phi_0-\phi_l)+\frac{A_u}{2A_0}\cos(\phi_u-\phi_0)\right\}.\,m\cos\omega t\right.$$
$$\left.-\left\{\frac{A_l}{2A_0}\sin(\phi_0-\phi_l)+\frac{A_u}{2A_0}\sin(\phi_u-\phi_0)\right\}.\,m\sin\omega t\right] \qquad (288)$$

and that for the [Y] component:

$$[Y]=A_0\left[\left\{\frac{A_l}{2A_0}\sin(\phi_0-\phi_l)-\frac{A_u}{2A_0}\sin(\phi_u-\phi_0)\right\}.\,m\cos\omega t\right.$$
$$\left.+\left\{\frac{A_l}{2A_0}\cos(\phi_0-\phi_l)-\frac{A_u}{2A_0}\cos(\phi_u-\phi_0)\right\}.\,m\sin\omega t\right] \qquad (289)$$

It should be noted that the steady carrier level corresponding to the constant term $A_0$ in equation 274 is omitted in the expression 288 for [X], since, when $\omega\rightarrow 0$, $A_l=A_u=A_0$ and $\cos(\phi_0-\phi_l)=\cos(\phi_u-\phi_0)=1$.

Comparison of these [X] and [Y] components with the original envelope (equation 271) shows that each is a distorted version of this envelope. Then the [X] component contains a term ($m\cos\omega t$), the modulating function, but to this is added a term ($m\sin\omega t$). Similarly, the [Y] component contains terms in both ($m\cos\omega t$) and ($m\sin\omega t$). Then each component may be considered to arise from the application of the envelope wave to two complex admittances—one for each of the [X] and [Y] components. Suppose we write these admittances as $Y_x(\omega)$ and $Y_y(\omega)$ respectively, and the admittance of the complete asymmetric sideband channel as $Y(\omega)$ (the admittance represented by the characteristics in Fig. 95). If also

we write the modulating envelope, for simplicity, as $F(\omega t)$, the modulated carrier wave applied to the input terminals of the channel is, from equation 271:

$$e_1 = E[F(\omega t)] \,.\, \cos \omega_0 t \quad \ldots\ldots \quad (290)$$

Then the response signal at the output terminals of the channel is, from 275:

$$\begin{aligned} e_1 \,.\, Y(\omega) = {} & EA_0[F(\omega t) \,.\, Y_x(\omega)] \cos(\omega_0 t + \phi_0) \\ & + EA_0[F(\omega t) \,.\, Y_y(\omega)] \sin(\omega_0 t + \phi_0) \quad \ldots\ldots \quad (291) \end{aligned}$$

But we have seen, from 288 and 289, that each of these admittances, $Y_x(\omega)$ and $Y_y(\omega)$, must be *complex*:

Let
$$\left.\begin{aligned} Y_x(\omega) &= g_x(\omega) + jb_x(\omega) \\ Y_y(\omega) &= g_y(\omega) - jb_y(\omega) \end{aligned}\right\} \quad \ldots\ldots \quad (292)$$

so that the response wave becomes:

$$\begin{aligned} e_1 \,.\, Y(\omega) = {} & EA_0[F(\omega t)\{g_x(\omega) + jb_x(\omega)\}] \cos(\omega_0 t + \phi_0) \\ & + EA_0[F(\omega t)\{g_y(\omega) - jb_y(\omega)\}] \sin(\omega_0 t + \phi_0) \quad \ldots \quad (293) \end{aligned}$$

The terms $g_x(\omega)$, $b_x(\omega)$, $g_y(\omega)$ and $b_y(\omega)$ are contained in { } in equations 288 and 289:

$$\left.\begin{aligned} g_x(\omega) &= \frac{A_l}{2A_0} \cos(\phi_0 - \phi_l) + \frac{A_u}{2A_0} \cos(\phi_u - \phi_0) \\ b_x(\omega) &= \frac{A_l}{2A_0} \sin(\phi_0 - \phi_l) + \frac{A_u}{2A_0} \sin(\phi_u - \phi_0) \\ g_y(\omega) &= \frac{A_l}{2A_0} \sin(\phi_0 - \phi_l) - \frac{A_u}{2A_0} \sin(\phi_u - \phi_0) \\ b_y(\omega) &= \frac{A_l}{2A_0} \cos(\phi_0 - \phi_l) - \frac{A_u}{2A_0} \cos(\phi_u - \phi_0) \end{aligned}\right\}^{*} \quad (294)$$

The complex admittances { } in equation 293 are the equivalent modulation-frequency characteristics and it is seen that *two* sets are necessary. It is to these admittances in turn that we apply the envelope wave $F(\omega t)$ in order to calculate the in-phase and quadrature components $[X]$ and $[Y]$ which, when added orthogonally, give the complete distorted envelope according to equation 277.

Let us consider now the form of these equivalent modulation-frequency characteristics in relation to the type of characteristics that we used originally, Fig. 95. These characteristics have been replotted in Fig. 102 (a) showing the carrier tuned to the *mid-band* frequency (the symmetric-sideband case) together with the equiva-

* See footnote on p. 227.

lent modulation-frequency characteristics obtained, in this simple case, merely by transferring the carrier to zero frequency. This equivalent characteristic is a $Y_x(\omega)$ component, since $Y_y(\omega)=0$ in this symmetric-sideband case (as seen by putting $A_l=A_u$ and $\phi_0-\phi_l=\phi_u-\phi_0$). Fig. 102 (b) shows the asymmetric sideband tuning, together with the *two* resulting equivalent modulation-frequency characteristics $Y_x(\omega)$ and $Y_y(\omega)$. These characteristics

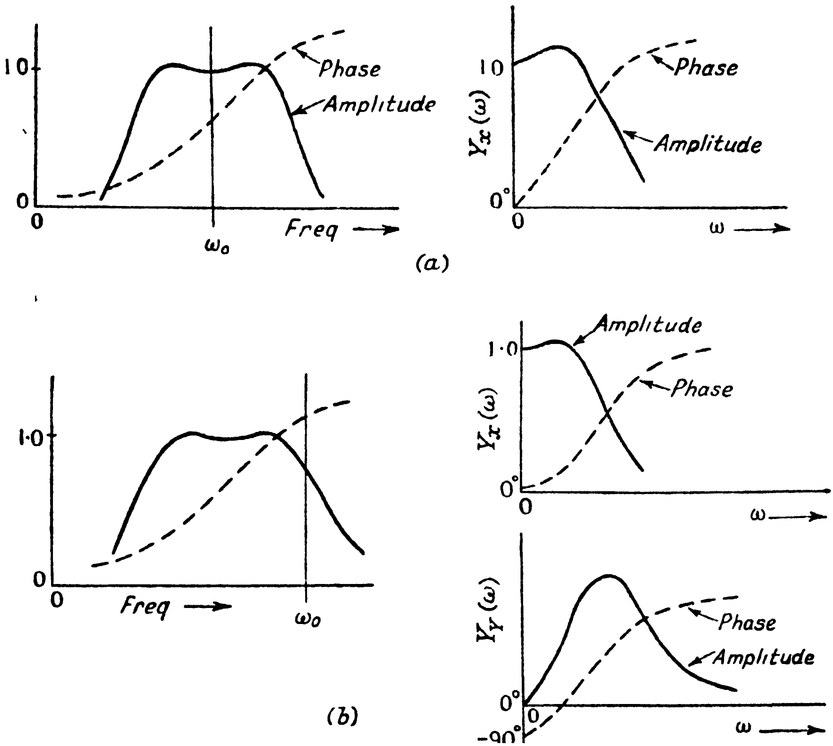



Fig. 102.—Bandpass and Equivalent Modulation-frequency Characteristics.
(a) Symmetrical case. (b) Asymmetrical case.

have been plotted in terms of modulus and phase shift, whereas equation 294 gives them in the form of their real and imaginary parts. We may readily obtain one from the other:

$$\left.\begin{array}{ll} Y_x(\omega) & \begin{cases} \text{Modulus} = \sqrt{(g_x{}^2(\omega)+b_x{}^2(\omega))} \\ \text{Phase shift} = \tan^{-1} b_x(\omega)/g_x(\omega) \end{cases} \\ Y_y(\omega) & \begin{cases} \text{Modulus} = \sqrt{(g_y{}^2(\omega)+b_y{}^2(\omega))} \\ \text{Phase shift} = -\tan^{-1} b_y(\omega)/g_y(\omega) \end{cases} \end{array}\right\} \quad . \quad . \quad (295)$$

where $g_x(\omega)$, $b_x(\omega)$ . . . , etc., are related to the channel characteristics as in equation 294.

The following points should be noted (see equations 292 and 294) in connection with these two sets of equivalent modulation-frequency characteristics:

$$\left.\begin{array}{r} |Y_x(\omega)|=1 \\ \text{Phase shift of } Y_x(\omega)=0 \\ |Y_y(\omega)|=0 \\ \text{Phase shift of } Y_y(\omega)=-90^\circ \end{array}\right\} \text{at } \omega=0 \quad . \quad . \quad (296)$$

since, when the modulation frequency $\omega=0$, we have $A_l=A_u=A_0$ and $\phi_l=\phi_u=\phi_0$.

The phase shift of $-90^\circ$ at zero frequency for the $Y_y(\omega)$ component is shown in Fig. 102 (b).

So far we have proved only that these equivalent modulation-frequency characteristics apply to pure sinusoidal modulation envelopes, but it will be shown later (Sec. 65) that they apply to any envelope signal waveform whatsoever.

## 63. Example: mistuning of a tuned amplifier

As a practical example of an asymmetric sideband channel let us consider an amplifier stage, consisting of a tetrode valve loaded by a tuned circuit, such as may be used for a preamplifier in a radio receiver. Such a circuit may be mistuned with respect to the incoming carrier frequency, so we may consider this mistuning to be of varying amounts and thereby calculate the equivalent modulation-frequency characteristics $Y_x(\omega)$ and $Y_y(\omega)$ in various cases.

Fig. 103 shows (insert) such an amplifier stage, together with its frequency-response characteristics, which have been drawn in "universal" form. We have already shown the characteristics of such a tuned circuit in Chapter 3, but we were dealing there with very low Q values in order to show the inherent asymmetry of the characteristics. The curves shown in Fig. 103 are adapted from those of Terman [12] and apply to reasonably high Q values, say 25–300, within an error of 1 per cent., such as may be used for an R.F. amplifier, and assume that the Q value remains constant over the entire band. Such characteristics are quite symmetrical about the mid-band frequency and give no asymmetric sideband distortion to a signal correctly tuned.

If $g$ is the mutual conductance of the valve (a tetrode type), then the valve anode current is given by $i_a=ge_1$ and the transfer characteristics of such a stage may be expressed in terms of the ratio of

the output to the input voltages, $e_2/e_1$. The output voltage of the stage is $e_2=i_a \,.\, Z(\omega)$, where $Z(\omega)$ is the impedance of the tuned-anode load. Thus the transfer characteristics are given by:

$$e_2/e_1=g \,.\, Z(\omega) \quad . \quad . \quad . \quad . \quad . \quad . \quad (297)$$

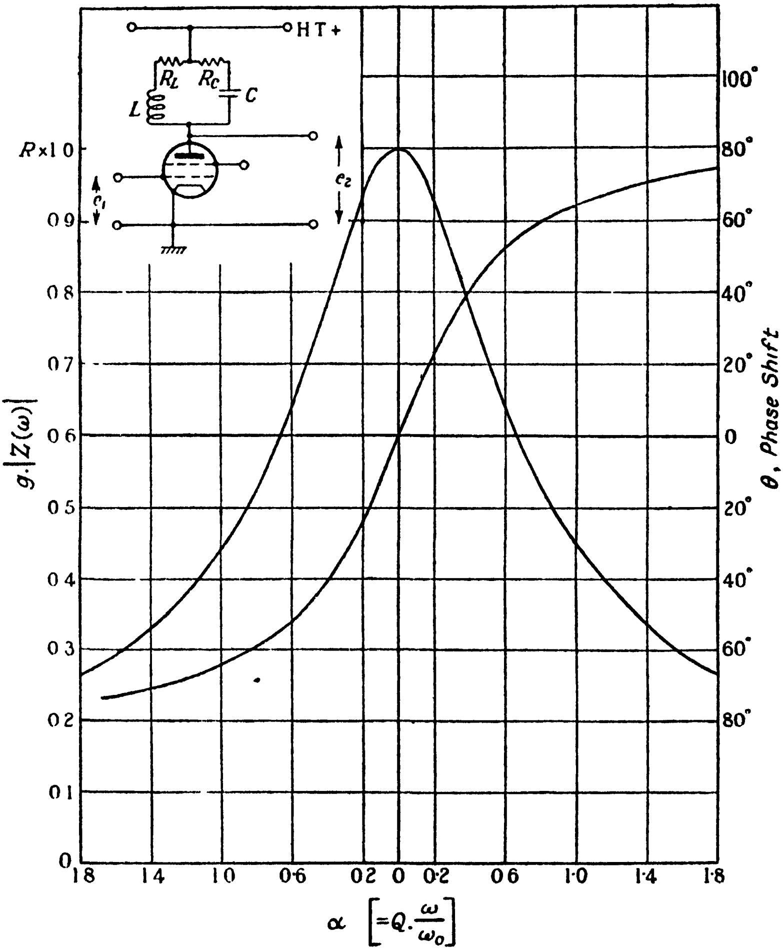
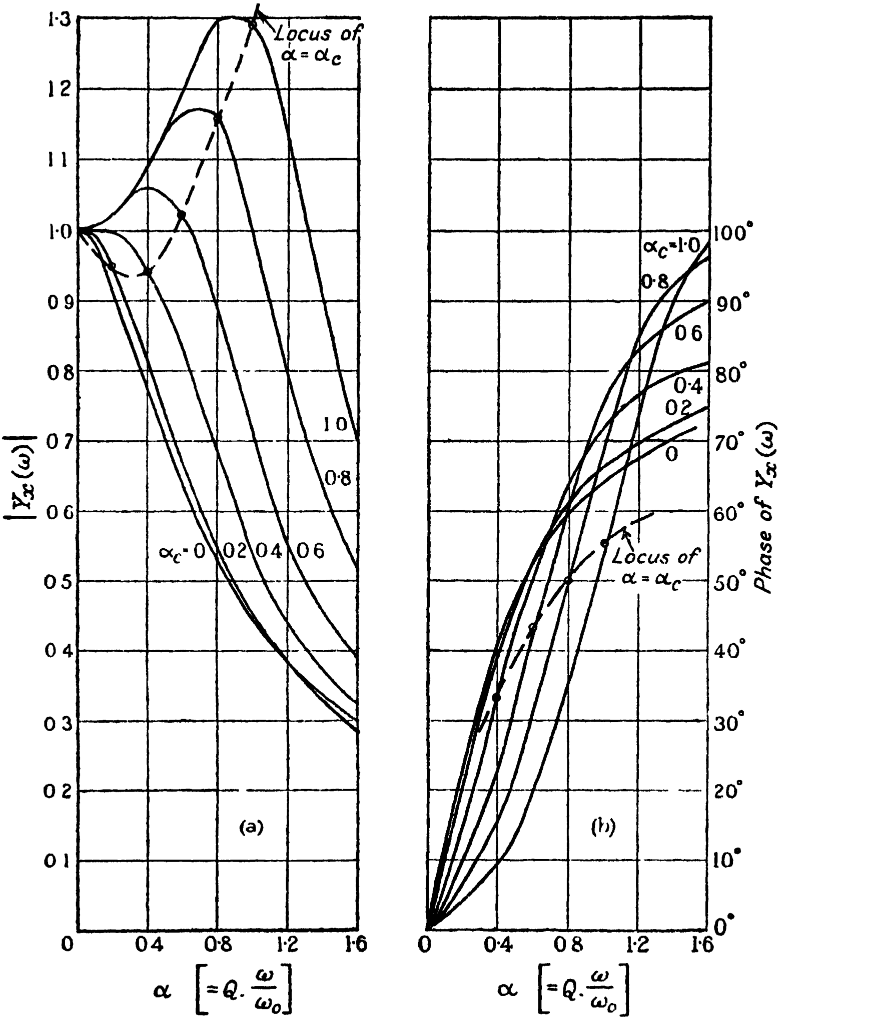



Fig. 103.—Frequency Characteristics of a Tuned Amplifier Stage.

so that they are proportional to $Z(\omega)$. The curves of the modulus and phase-shift of $e_2/e_1$ are shown in Fig. 103, plotted against a scale of $\alpha$ ($=Q \,.\, \omega/\omega_0$), $\omega_0$ being the mid-band frequency in this symmetrical case.

Now consider the carrier to be tuned off the mid-band frequency by various amounts; $\alpha_c$ can represent the particular value of $\alpha$ on the universal curve, Fig. 103, to which the carrier is tuned in each



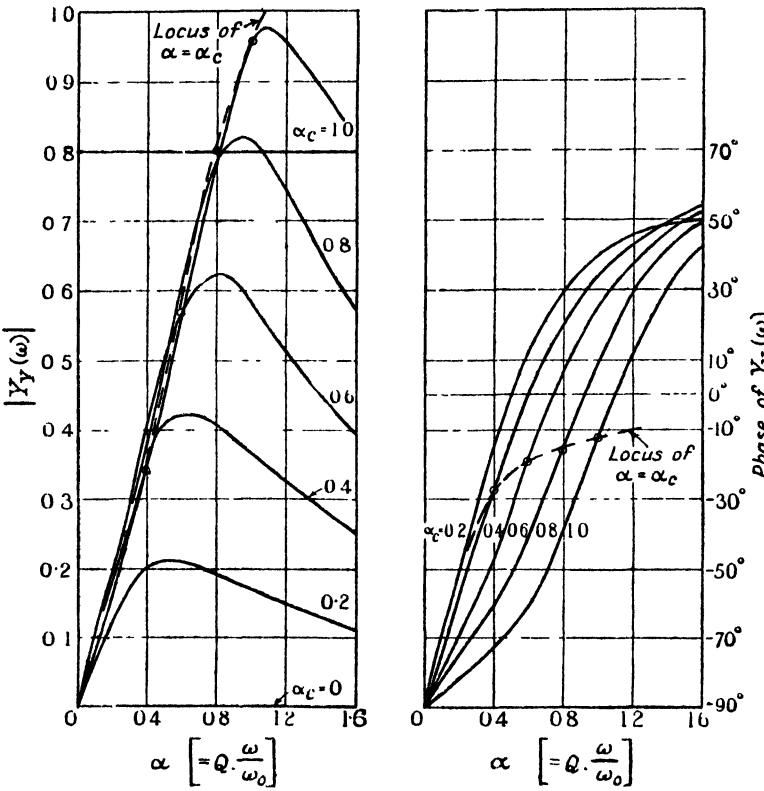

(a) $Yx(\omega)$

Fig. 104.—Amplitude and Phase-shift Components of $Y_x(\omega)$ and $Y_y(\omega)$ for the Tuned Amplifier (Fig. 103) with Mis-tuned Carrier.

case. The values of $Y_x(\omega)$ and $Y_y(\omega)$, the equivalent modulation-frequency characteristics, have been calculated from these universal modulus and phase-shift curves, using the equations 294 and 295,

for tuning positions of the carrier frequency, $\omega_0$, corresponding to $\alpha_c = 0{\cdot}2$, $0{\cdot}4$, $0{\cdot}6$, $0{\cdot}8$ and $1{\cdot}0$. These values are plotted in Fig. 104. Naturally the $Y_x(\omega)$ curves, corresponding to $\alpha_c = 0$, are identical with the original curves in Fig. 103 (since these represent the case of the carrier tuned to mid-band) with the mid-band frequency transferred to zero. Also, the curves for $Y_y(\omega)$ vanish for this case.

It can be seen from Fig. 104 that as the carrier frequency departs more and more from the mid-band frequency ($\alpha_c$ increasing) the magnitude of the admittance $Y_y(\omega)$ becomes more appreciable, particularly towards the higher values of $\omega$ (the modulation frequency).




(b) $Y_y(\omega)$

Fig. 104 (continued).

Also as $\alpha_c$ is increased, the peak in the modulus of both $Y_x(\omega)$ and $Y_y(\omega)$ increases in frequency and amplitude. Thus if we consider the distortion of a modulation envelope, as given by these

admittances, it would appear that at certain frequencies, given by the peaks, pronounced overswings or ripples might occur. The magnitude and duration of these overswings would depend on the sharpness of these peaks, a matter which has already been discussed in Sec. 37. It might be expected from a superficial examination of the frequency characteristics in Fig. 103 that if the carrier be mistuned from mid-band, to a frequency corresponding to $\alpha_c$, then the distorted response envelope would contain overswings at this frequency, or near to it, since the sidebands at this frequency would be accentuated. Actually the peaks in $Y_x(\omega)$ and $Y_y(\omega)$ do not occur at exactly the frequency corresponding to $\alpha_c$, the mistune frequency, and the dotted curves on Figs. 104 (a) and (b) give the locus of $\alpha = \alpha_c$, showing the slight displacement of the peaks.

With regard to the phase shifts of $Y_x(\omega)$ and $Y_y(\omega)$ it can be seen that they are both non-uniform, so that both the $[X]$ and $[Y]$ components of the response envelope wave will possess phase distortion, which must introduce an element of asymmetry in their waveforms (see Sec. 39). Furthermore, the slopes of both these phase-shift characteristics decrease as $\alpha$ increases, for low values of $\alpha_c$, whereas the slopes increase as $\alpha$ increases for higher values of $\alpha_c$ (carrier tuned farther away from resonance). Thus, as the carrier is tuned farther and farther off mid-band, the higher frequency components in the response envelope will be delayed more and more with respect to the low-frequency components (see Sec. 40 on phase delay). The effect of this varied delay on an applied square wave signal (a picture element) or a pulse is to produce a small burst of energy just after the main signal.[13, 4]

The increase in the delay of the higher modulation frequencies, when the carrier is tuned well away from the mid-band, is not typical only of tuned circuit amplifier stages, but of any band-pass channel which has the same type of phase-shift characteristic—that is, of a certain slope and reasonably linear over the pass-band, but curving off to become horizontal in the region of cut-off.

## 64. Modulation by a square-wave envelope

We have so far dealt only with a carrier wave modulated by a pure tone, having a sinusoidal envelope. This simple case shows most clearly how distortion arises in single- or asymmetric sideband channels, with its dependence on modulation depth, and also the interrelation between the resulting envelope distortion and the phase modulation of the carrier. It also enables us to develop a

vectorial means of examining the problem, so that by dividing the distorted response wave into two orthogonal components, [X] and [Y], a set of complex admittances may be obtained, $Y_x(\omega)$ and $Y_y(\omega)$, which enable us to calculate the waveform distortion in any particular case by transferring the carrier to zero frequency.

Pure sinusoidal modulation is of limited interest, so let us now extend the results we have obtained to other forms of envelope. The square-wave or step-wave signal is perhaps of greatest importance, and the distortions of such envelope waveforms should be examined in detail. Such signals may be regarded as limiting "picture elements" in television; also the results obtained may show the type of distortion to be expected with rectangular pulse signals.

For detailed examination let us take the periodic square-wave envelope representable by a Fourier series of harmonics. The results will be approximately correct for a single step-wave or a rectangular pulse, at least in a qualitative way, although such waveforms involve a continuous spectrum. We have already considered, in Sec. 37, the approximation to a continuous spectrum by the use of a series of harmonic terms and have calculated (Chapter 4, Sec. 35) the harmonic series for a square wave, illustrated by Fig. 25. Modulation of a carrier wave of amplitude $E$ to a depth of $100m$ per cent. means that the envelope amplitude is $\pm mE$. The modulated wave may then be written:

$$e_1 = E\left[1 + \frac{4m}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\frac{2n\pi t}{T_0}\right] . \cos\omega_0 t \quad . \quad . \quad (298)$$

where $n$ is odd (odd harmonics only). $T_0$ is the repetition period of the square-wave envelope. This expands into a series of upper and lower sidebands:

$$e_1 = E\cos\omega_0 t + E\,.\,\frac{2m}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\left(\omega_0 + \frac{2n\pi}{T_0}\right)t$$

$$-E\,.\,\frac{2m}{\pi}\sum^{\infty}\frac{1}{n}\sin\left(\omega_0 - \frac{2n\pi}{T_0}\right)t \quad . \quad . \quad . \quad . \quad (299)$$

Let this wave be applied to a band-pass filter, asymmetrically, with the carrier frequency $\omega_0$ tuned to one side as illustrated in Fig. 95. Again let $A_u$, $A_l$, $\phi_u$ and $\phi_l$ represent the modulus and phase shift of the filter at the various upper and lower sideband

frequencies; these are really functions of $n$ for any given sideband.

The wave at the output terminals of the filter is then:

$$e = EA_0 \cos(\omega_0 t + \phi_0) + E \cdot \frac{2m}{\pi} \sum_{n=1}^{\infty} \frac{A_u}{n} \sin\left[\left(\omega_0 + \frac{2n\pi}{T_0}\right)t + \phi_u\right]$$

$$- E \cdot \frac{2m}{\pi} \sum_{n=1}^{\infty} \frac{A_l}{n} \sin\left[\left(\omega_0 - \frac{2n\pi}{T_0}\right)t + \phi_l\right] \quad (300)$$

As in the case of pure sine-wave modulation, described in Sec. 60, so in this case of modulation by a square-wave envelope, we may represent the modulated carrier by a vector diagram both before

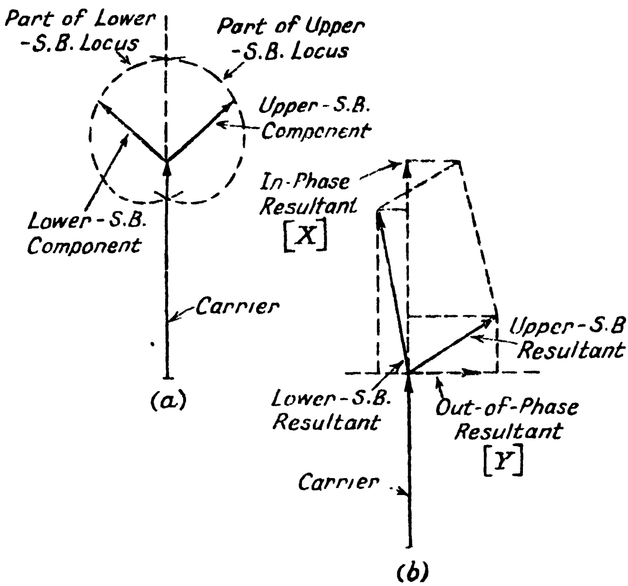



Fig. 105.—Vector Diagram of Modulated Wave (Complex Modulation).

and after its passage through the asymmetric sideband filter. We may regard the carrier vector as stationary on the paper and each modulation component, of any particular modulation frequency $(2n\pi/T_0)$, as consisting of two sideband vectors symmetrically disposed on either side of the carrier vector but rotating in opposite directions. If every modulation frequency is now included we have a number, possibly infinite, of pairs of rotating sideband vectors (see Fig. 105 (a)).

The sum of all the sideband vectors at any instant of time has a component in phase with the carrier which is equal to the envelope amplitude at that instant, while the sum of the components in quadrature with the carrier must be zero. The resultant of all the upper sidebands is a vector which varies in length and angular velocity, tracing out a certain locus which depends on the form of the carrier envelope; similarly the resultant of the lower sidebands is the mirror image.

When the wave is passed through the asymmetric sideband filter, so that the sidebands undergo amplitude and phase distortion, the resultant upper and lower sidebands may no longer be symmetrically disposed about the carrier vector and so the resultant sideband vector component in quadrature with the carrier is no longer zero (Fig. 105 (b)). This quadrature component $[Y]$ introduces distortion into the wave envelope and also carrier phase-modulation, as we have already seen in the case of pure sine-wave modulation. The in-phase $[X]$ and quadrature $[Y]$ components may be found from the equation for the output wave (300) by expanding this into:

$$e=\left[A_0+\frac{2m}{\pi}\sum_{n=1}^{\infty}\frac{A_u}{n}\sin\left(\frac{2n\pi t}{T_0}+\phi_u-\phi_0\right)\right.$$
$$\left.+\frac{A_l}{n}\sin\left(\frac{2n\pi t}{T_0}+\phi_0-\phi_l\right)\right].E\cos(\omega_0 t+\phi_0)$$
$$+\left[\frac{2m}{\pi}\sum_{n=1}^{\infty}\frac{A_u}{n}\cos\left(\frac{2n\pi t}{T_0}+\phi_u-\phi_0\right)\right.$$
$$\left.-\frac{A_l}{n}\cos\left(\frac{2n\pi t}{T_0}+\phi_0-\phi_l\right)\right].E\sin(\omega_0 t+\phi_0) \quad . \quad . \quad (301)$$

which may be written:

$$e=[X].E\cos(\omega_0 t+\phi_0)+[Y].E\sin(\omega_0 t+\phi_0) \quad . \quad . \quad (302)$$

As in the case of pure sine-wave modulation, the distorted envelope and carrier phase-modulation are given by $[X]$ and $[Y]$ (see equations 277 and 278).

Comparison of the expression 301 with that for the input square-wave modulated carrier (298) shows once again that $[X]$, the in-phase component, bears some resemblance to the original envelope, though the various sideband components are altered in amplitude and phase. The quadrature component $[Y]$ arises owing to the

sideband asymmetry, and introduces distortion. This component is seen to be proportional in magnitude to $m$, the modulation coefficient, so, in this example also, the asymmetric sideband channel distortion increases with the depth of modulation.

The distortion of the $[X]$ component arises from two sources, the non-uniform amplitude variations and phase-shifts of the various components in the envelope. As is true of the frequency distortion of any linear network, the effect of a non-uniform phase shift characteristic is to give a non-symmetrical appearance to the response (with a symmetrical applied waveform*). Thus if we compare the $[X]$ component in equation 301 with the original square-wave envelope (see 298) we see that every sine component is split into two components differing in phase. However, if we assume that the phase-shift characteristic of the channel is straight and that only the modulus characteristic produces sideband asymmetry, we may write this phase shift as:

$$\left.\begin{aligned} \phi_u &= \phi_0 + \frac{2n\pi t_1}{T_0} \\ \phi_l &= \phi_0 - \frac{2n\pi t_1}{T_0} \end{aligned}\right\} \quad . \quad . \quad . \quad . \qquad (303)$$

where $t_1$ is the slope† of the characteristic, $d\phi/d\omega$. The shift in phase, relative to the carrier phase $\phi_0$, of any sideband is now proportional to its frequency. Substituting this condition in the $[X]$ component of expression 301 gives:

$$[X] = A_0 + \frac{2m}{\pi} \sum_{n=1} \frac{(A_u + A_l)}{n} \sin \frac{2n\pi}{T_0}(t + t_1) \quad . \quad (304)$$

which is similar to the original envelope (see 298) except that it is delayed by a time interval $t_1$ and that the various components are modified in their amplitudes, producing a symmetrical distortion.

The quadrature component $[Y]$ becomes, similarly:

$$[Y] = \frac{2m}{\pi} \sum_{n=1}^{\infty} \frac{(A_u - A_l)}{n} \cos \frac{2n\pi}{T_0}(t + t_1) \quad . \quad . \quad (305)$$

The actual shape of the response envelope waveform depends in detail upon the values of $A_u$ and $A_l$ for the various sideband com-

* See Sec. 39.
† See Sec. 40.

ponents, but there are certain forms of distortion which are typical of asymmetric sideband channels and which are most serious in this case of square-wave modulation. Step-wave or pulse modulation, or any modulation envelope which possesses discontinuity, will also exhibit the same type of distortion, but the square wave, with its limited series of harmonic components, presents the simplest calculation. The following examples illustrate the nature of this distortion.

Consider a carrier wave to be modulated by a square-wave envelope to a depth of 100 per cent. ($m=1 \cdot 0$) and tuned to one side of a band-pass channel, as in Fig. 95. This full modulation depth is chosen as giving the maximum distortion. Assume the phase-shift characteristic to be of uniform slope (as 303) and let the modulus of the channel characteristics give the following values of $A_u$ and $A_l$:

| Harmonic number, $n$ | $A_u$ | $A_l$ |
|---|---|---|
| 0 (carrier) | $A_0=0 \cdot 5$ | |
| 1 | 0·3 | 0·705 |
| 3 | 0·088 | 0·862 |
| 5 | 0·013 | 0·888 |
| 7 | 0 | 0·89 |
| 9 | 0 | 0·84 |
| 11 | 0 | 0·53 |
| 13 | 0 | 0·09 |

The channel is thus wide enough to pass effectively thirteen sideband components; $n$, of course, has only odd values since the even harmonics in the square-wave envelope are zero. The carrier here has been assumed to be tuned half-way down the side of the channel characteristic, so that $A_0=0 \cdot 5$. The magnitudes of the $[X]$ and $[Y]$ components are found by substituting these figures in expressions 304 and 305, since the phase shift is uniform. These two components are plotted orthogonally, in Fig. 106, so as to give the vector locus diagram for this example. In this figure the ringed points indicate the positions of the resultant vector at equal increments* of time $t=T_0/18$. Both the angular velocity of the vector and its length vary with time, but at any instant the angular position $\psi$ gives the instantaneous carrier phase shift, while the vector length gives

* The vector is stationary for a while at both the points, on the locus, corresponding to $Y=0$, so that some ringed points are coincident there.

the envelope amplitude. The waveforms of [$X$] and [$Y$] are plotted in Fig. 107 (dotted lines) together with the distorted envelope wave-

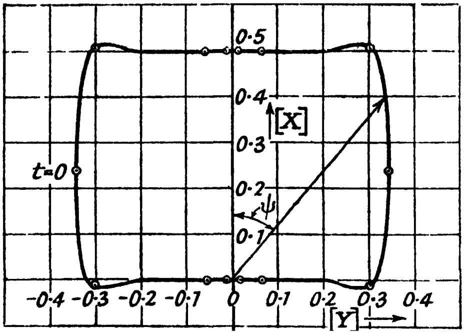



Fig. 106.—Square-wave Modulation—Vector Locus.

form (full line) $\sqrt{(X^2+Y^2)}$ showing the effects of the sideband asymmetry. It may be seen that the [$X$] component is similar to the original square-wave envelope, except that its build-up time is finite,

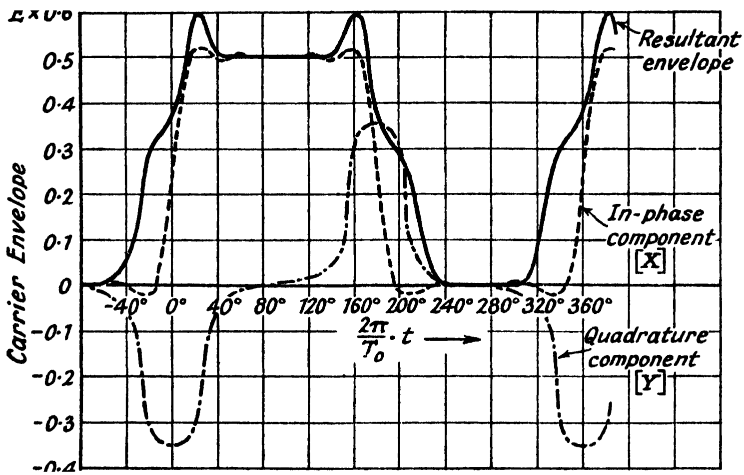



Fig. 107.—Square-wave Envelope—Distortion with Practical Asymmetric-sideband Filter.

as would be expected from the restricted band-width of the channel. The [$Y$] component adds distortion, so that the resulting envelope shows sharp "spikes" at points corresponding to the edges of the

original square waves—the points of discontinuity—also the build-up time is deteriorated and the mark/space ratio of the signal is increased.

These are the most serious effects on the signal transmission and are typical of all asymmetric sideband channels. It is true that even a normal double-sideband channel of restricted band-width causes overswings at the points of discontinuity of such a signal, as we have seen from examples in Chapter 5. These overswings do not, however, reach the proportions of the "spikes" that can arise due to asymmetric sideband distortion.

However, the usefulness of asymmetric sideband channels for television should not be dismissed on the grounds of such distortion, since in practice its effects on the signal quality may not be as serious as might appear at first.[4] For one thing, we have considered 100 per cent. modulation here, giving the maximum distortion, whereas the trough of the envelope never reaches lower than 25 per cent. of the peak carrier amplitude in the British system of television,* the remainder being used for the synchronising signal. Also a certain degree of overswing on the sharp edges of signal elements is sometimes desirable (see footnote on p. 177).

The overswing that arises in ordinary double-sideband channels of restricted band-width has been illustrated in Chapter 5, Fig. 72, using an idealised "flat-topped" filter-response curve. Let us now see what distortion is set up when the carrier is tuned to one edge of such an idealised channel.

The inset diagram in Fig. 108 shows the modulus of the idealised "flat-topped" characteristics; the phase-shift characteristic is assumed to be of uniform slope so that its effects may be neglected, a mere time-delay being of no consequence at present. We have already discussed the non-physical nature of such characteristics, and the justification of their use, in Chapter 5. The carrier, of angular frequency $\omega_0/2\pi$, is shown tuned at the upper cut-off edge of this filter so that the upper sideband is completely removed and a band-width has been chosen that just accommodates the lower sidebands up to a frequency $13/T_0$ so as to line up with the practical example we have already considered. Then $A_l = 1{\cdot}0$ for $n \leqslant 13$, while $A_u = 0$ always. Let $A_0 = 0{\cdot}5$, as though the carrier were tuned half-way down the side of the filter characteristic (it may be assumed that the edge of the filter characteristic slopes just perceptibly).

* It reaches to zero carrier level in the American system.

Substitution of these values in equations 304 and 305 gives:

$$\left.\begin{aligned}[X] &= 0{\cdot}5 + \frac{2}{\pi}\sum_{n=1}^{13}\frac{1}{n}\sin\frac{2n\pi}{T_0}\,.\,t \\ [Y] &= \quad -\frac{2}{\pi}\sum_{n=1}^{13}\frac{1}{n}\cos\frac{2n\pi}{T_0}\,.\,t\end{aligned}\right\} \quad . \; . \; . \; . \quad (306)$$

These components are plotted (dotted lines) in Fig. 108, together with the resultant envelope waveform, $\sqrt{(X^2+Y^2)}$. Once again

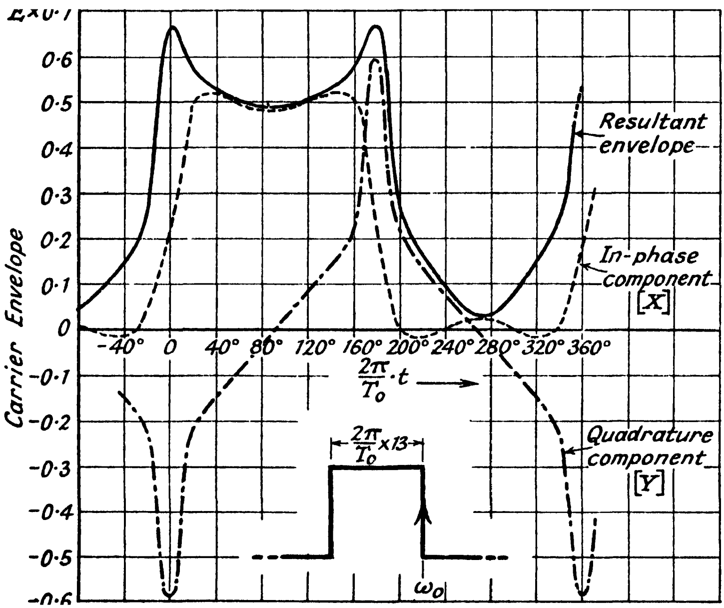



Fig. 108.—Square-wave Envelope—Idealized Single-sideband Distortion.

the $[X]$ component closely approaches the original square envelope in shape, while the $[Y]$ component introduces distortion. The distortion again takes the form of sharp "spikes" at the edges of the square envelope combined with a deterioration of the build-up time. It is interesting to compare this envelope shape with the waveform that results from tuning the carrier to the mid-band of the idealised filter, shown in Fig. 72. The reduction of the build-up rate of the wave is most marked at the troughs of the envelope, corresponding to high depths of modulation, and in this extreme case, where

$m=1{\cdot}0$, the troughs between successive square waves are virtually filled up. The effect of such a shape of build-up curve is an apparent reduction in the modulation depth, but this serious form of distortion rapidly reduces as $m$ is decreased.

This type of distortion, which is set up in asymmetric sideband channels, makes such channels rather unsuitable for pulse communication systems, at least if the received pulse width is of importance. For instance, in certain radar systems the time of arrival of the leading edge of a pulse determines the accuracy of operation, so that the effective widening at the bottom of the pulse and the spoilation of the build-up rate there can deteriorate this accuracy. Better results may possibly be achieved by working double-sideband with a slightly reduced band-width; in this way the build-up rate is somewhat reduced, but it is more uniform from top to bottom of the pulse and is not made markedly worse at the bottom. At the same time the peak signal-to-noise ratio is correspondingly improved by the band-width reduction.

Characteristics of this idealised flat-topped form have been adopted by several investigators [2, 3, 7] for the purpose of asymmetric sideband distortion analysis, but the results can be very pessimistic, giving a distortion greater than would ever arise in practical networks owing to the extreme rate of cut-off in the region of the carrier frequency. For example, the response, Fig. 107, of the practical filter shows far less distortion than does the response of the idealised filter, Fig. 108. The "spikes" that arise at the edges of the envelope may be considered as a high-frequency distortion, while the reduction of the build-up rate (usually an effect of high-frequency cut-off in symmetric channels) is due partly to low-frequency distortion. It is true that the $[X]$ component has itself a finite build-up time due only to the finite band-width, but the $[Y]$ component introduces not only the sharp "spikes" but also a strong low-frequency ripple of period $T_0$. This low-frequency ripple is most marked in the idealised case (Fig. 108) and is the cause of the filling-in of the troughs of the response envelope waveform.

This last element of distortion, being a low-frequency effect, must be produced by the sidebands closest to the carrier. In the case of the practical filter the carrier is tuned on one side of the characteristic which has a finite cut-off rate, so that both $A_u$ and $A_l$ are finite for the innermost sidebands (the "vestigial" sidebands as they are sometimes called). In this case $(A_l-A_u)$ can be negligibly small for a range of low frequencies, so that, as may be seen from equation

305, the [$Y$] component may have negligible low-frequency components. In the idealised case, however, $(A_l - A_u)$ is finite at all frequencies from zero to cut-off, so as to introduce low-frequency distortion of a serious magnitude.

## 65. Transient modulation—equivalent modulation-frequency characteristics

It has been shown already (Sec. 62) that, in the case of modulation by a pure sinusoidal envelope, the carrier frequency may effectively be reduced to zero, so that a set of equivalent modulation-frequency characterictics can be derived from which the envelope distortion may be determined directly. It was also shown that *two* sets of such characteristics are required, denoted by $Y_x(\omega)$ and $Y_y(\omega)$ for asymmetric sideband channels.

Let us now see whether these same characteristics, given by equations 292 and 294, are applicable to modulation of other forms, using a general transient envelope for the purpose. We can write the amplitude-modulated wave, in this case, as:

$$e_1 = E\left[1 + \int_0^\infty \{a(\omega)\cos\omega t + b(\omega)\sin\omega t\}d\omega\right] . \cos\omega_0 t \qquad (307)$$

The expression in the square brackets represents the envelope, consisting of a continuous spectrum of both cosine, $a(\omega)$, terms and sine, $b(\omega)$, terms, together with a D.C. component. Such continuous transient spectra have been discussed in detail in Sec. 21. Chapter 2.

Expanding the above modulated-wave expression into its carrier, upper and lower sidebands, gives:

$$e_1 = E\cos\omega_0 t + \frac{E}{2}\int_0^\infty a(\omega)\cos(\omega_0+\omega)t + b(\omega)\sin(\omega_0+\omega)t.d\omega$$
$$+\frac{E}{2}\int_0^\infty a(\omega)\cos(\omega_0-\omega)t - b(\omega)\sin(\omega_0-\omega)t.d\omega \quad . \qquad (308)$$

Applying this wave to a band-pass channel, with the carrier tuned to one side of the pass-range, as in Fig. 95, gives* for the distorted asymmetric sideband response wave:

$$e = EA_0\cos(\omega_0 t+\phi_0) + \frac{E}{2}\int_0^\infty A_u\left[a(\omega)\cos(\overline{\omega_0+\omega t}+\phi_u)\right.$$
$$\left. + b(\omega)\sin(\overline{\omega_0+\omega t}+\phi_u)\right]d\omega$$
$$+\frac{E}{2}\int_0^\infty A_l\left[a(\omega)\cos(\overline{\omega_0-\omega t}+\phi_l) - b(\omega)\sin(\overline{\omega_0-\omega t}+\phi_l)\right]d\omega \qquad (309)$$

* See footnote on p. 227.

This expression may be expanded and split into two orthogonal carrier components, just as in equation 274 for the pure tone modulation case; thus:

$$e=\Big[A_0+\tfrac{1}{2}\int_0^\infty A_l\{a(\omega)\cos(\omega t+\phi_0-\phi_l)+b(\omega)\sin(\omega t+\phi_0-\phi_l)\}$$
$$+A_u\{a(\omega)\cos(\omega t+\phi_u-\phi_0)$$
$$+b(\omega)\sin(\omega t+\phi_u-\phi_0)\}d\omega\Big].\cos(\omega_0 t+\phi_0)$$
$$+\Big[\tfrac{1}{2}\int_0^\infty A_l\{a(\omega)\sin(\omega t+\phi_0-\phi_l)-b(\omega)\cos(\omega t+\phi_0-\phi_l)\}$$
$$-A_u\{a(\omega)\sin(\omega t+\phi_u-\phi_0)$$
$$-b(\omega)\cos(\omega t+\phi_u-\phi_0)\}d\omega\Big].\sin(\omega_0 t+\phi_0)\,. \quad . \quad (310)$$

which, as before, may be written:

$$e=[X]\cos(\omega_0 t+\phi_0)+[Y]\sin(\omega_0 t+\phi_0) \quad . \quad (311)$$

where $[X]$ and $[Y]$ represent the components of the resulting distorted wave, respectively in phase and in quadrature with the carrier. There is no need to discuss further the significance of these two components; in this general case, as for the simple types of modulation considered previously, the vector locus of this distorted wave is given by these orthogonal components $[X]$ and $[Y]$, and they determine completely the envelope waveform and the carrier phase modulation, according to equations 277 and 278.

Also, as was the case with the pure tone modulation examined in Sec. 62, it can be seen that the $[X]$ and $[Y]$ components, as given by the expressions in square brackets in equation 310, are distorted versions of the original envelope. The various spectrum terms in this envelope are altered in amplitude and in phase, so that the original envelope, which may be written:

$$e_E=F(\omega t)=E[1+\int_0^\infty a(\omega)\cos\omega t+b(\omega)\sin\omega t\,.d\omega] \quad . \quad (312)$$

may be considered to have been applied to two complex admittances, the responses of which give the $[X]$ and $[Y]$ components.

Separation of the envelope terms and the filter characteristics $A_l$, $\phi_l$ . . . , etc., in the expression 310 for the distorted response wave shows that these complex admittances are $Y_x(\omega)$ and $Y_y(\omega)$, which are identical with those defined by equations 292 and 294 for

pure tone modulation, as may readily be proved by the reader by expansion of the expression 310. For example, we may write:

$$\cos(\omega t+\phi_0-\phi_l)=\cos\omega t\,.\,\cos(\phi_0-\phi_l)-\sin\omega t\,.\,\sin(\phi_0-\phi_l)$$

etc., and collect up the terms in $(\omega t)$.

There is no need, however, to go to this labour, since the proof follows directly from the proof for the pure tone (sinusoidal) modulation case. Both the admittances $Y_x(\omega)$ and $Y_y(\omega)$ are linear, and so, since they apply for any single modulating frequency, they must apply to any number of modulating terms appearing simultaneously. The fact that the resulting response signal from the asymmetric sideband channel varies in its distortion with the depth of modulation does not indicate that these equivalent modulation-frequency characteristics are themselves non-linear; each set of characteristics, $Y_x(\omega)$ and $Y_y(\omega)$ is itself linear, but the individual responses of each, $[X]$ and $[Y]$, vary in amplitude with the modulation depth, and these responses are then added *orthogonally* to give the resulting envelope.

The response wave (310) may thus be put in the form:

$$\begin{aligned} e=EA_0[F(\omega t)\{g_x(\omega)+jb_x(\omega)\}]\,.\,\cos(\omega_0 t+\phi_0)\\ +EA_0[F(\omega t)\{g_y(\omega)-jb_y(\omega)\}]\,.\,\sin(\omega_0 t+\phi_0) \quad . \quad . \quad [(293)]\end{aligned}$$

where the admittances $g_x(\omega)+jb_x(\omega)$ and $g_y(\omega)-jb_y(\omega)$ are defined by equation 294 in terms of the asymmetric sideband channel characteristics.

Thus, whatever the form of the envelope, $F(\omega t)$, of a modulated wave, the distortion due to asymmetric sideband operation may always be determined from this set of equivalent modulation-frequency characteristics. The carrier may effectively be transferred to zero frequency and the envelope $F(\omega t)$ may then be applied to each of the admittances $Y_x(\omega)$ and $Y_y(\omega)$ in turn, giving the orthogonal components $[X]$ and $[Y]$ of the resultant distorted wave. These components then determine the wave envelope and carrier phase modulation.

## 66. Simple method of obtaining these characteristics

A geometrical method of deriving these equivalent modulation-frequency characteristics has been pointed out by Nyquist [14] which depends on the fact that a set of asymmetrical characteristics, such as those in Fig. 95, may be split into two component sets, one symmetrical and the other skew-symmetrical about the carrier frequency. Before splitting the characteristics in this way, however, they must be put in their complex form and not left as modulus and

phase-shift characteristics. We shall now show that these component characteristics are identical with our $Y_x(\omega)$ and $Y_y(\omega)$ equivalent modulation-frequency characteristics.

Fig. 109 (a) shows a carrier, of frequency $\omega_0$, tuned to one side of a band-pass channel, as for asymmetric-sideband working. As before, let $A_0$ and $\phi_0$ be the modulus and phase-shift of the channel at the carrier frequency $\omega_0$ and let $A_u$, $A_l$, $\phi_u$, $\phi_l$ be the corresponding values for any pair of upper and lower sidebands. Then, taking $A_0\underline{|\phi_0}$ as a datum admittance, the lower sideband is modified, in

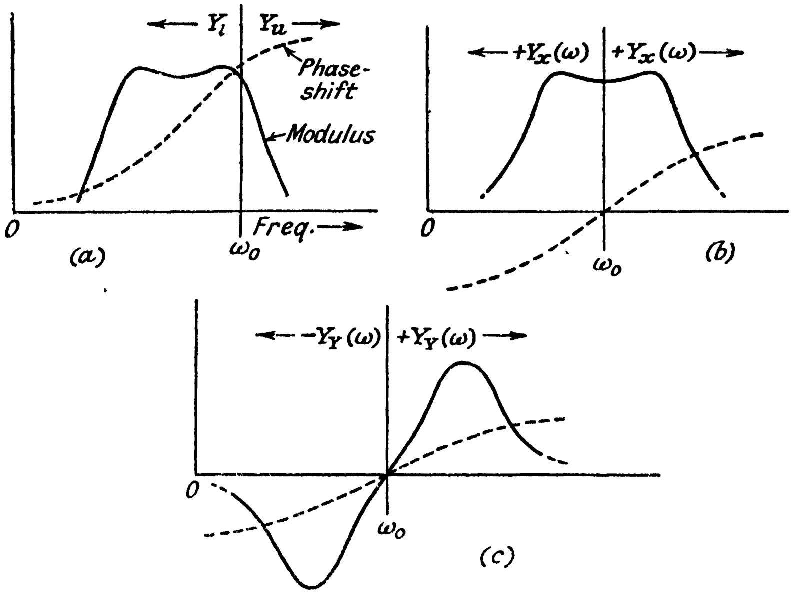



Fig. 109.—Division of Asymmetric Sideband Characteristics into Symmetrical and Skew-Symmetrical Components.

relation to the carrier, by the admittance $A_l/A_0\,\underline{|\phi_0-\phi_l}$, and the upper sideband is similarly modified by $A_u/A_0\,\underline{|\phi_u-\phi_0}$. Let us therefore make the following definitions:

$$\left.\begin{aligned} Y_l &= \frac{A_l}{A_0}\underline{|\phi_0-\phi_l} = \text{admittance* to lower sideband} \\ Y_u &= \frac{A_u}{A_0}\underline{|\phi_u-\phi_0} = \text{admittance* to upper sideband} \end{aligned}\right\} \quad . \quad (313)$$

* Again $Y_l$ and $Y_u$ are written short for $Y_l(\omega)$, $Y_u(\omega)$ and refer to any modulation frequency $\omega$.

If these are put in terms of their real and imaginary parts:

$$Y_l=\frac{A_l}{A_0}[\cos(\phi_0-\phi_l)+j\sin(\phi_0-\phi_l)]$$

$$Y_u=\frac{A_u}{A_0}[\cos(\phi_u-\phi_0)+j\sin(\phi_u-\phi_0)] \quad . \quad . \quad . \quad (314)$$

Now the asymmetric channel characteristics, illustrated by Fig. 109 (a), could be replotted in terms of $Y_l$ and $Y_u$ as given by this equation, 314. Since the carrier amplitude and phase shift, $A_0$, $\phi_0$, are taken as the datum here, the channel characteristics would then be plotted in relation to this datum and hence they are really expressed in terms of modulation frequency.

Adding and subtracting the admittances $Y_l$ and $Y_u$ in turn gives:

$$(Y_l+Y_u)=\left[\frac{A_l}{A_0}\cos(\phi_0-\phi_l)+\frac{A_u}{A_0}\cos(\phi_u-\phi_0)\right]$$
$$+j\left[\frac{A_l}{A_0}\sin(\phi_0-\phi_l)+\frac{A_u}{A_0}\sin(\phi_u-\phi_0)\right] \quad . \quad (315)$$

$$(Y_l-Y_u)=\left[\frac{A_l}{A_0}\cos(\phi_0-\phi_l)-\frac{A_u}{A_0}\cos(\phi_u-\phi_0)\right]$$
$$+j\left[\frac{A_l}{A_0}\sin(\phi_0-\phi_l)-\frac{A_u}{A_0}\sin(\phi_u-\phi_0)\right] \quad . \quad (316)$$

These admittances may, from equations 292 and 294, be seen to relate to $Y_x(\omega)$ and $Y_y(\omega)$, the equivalent modulation frequency characteristics, in the following way:

$$\left.\begin{aligned}\tfrac{1}{2}(Y_l+Y_u)&=Y_x(\omega)\\ \tfrac{1}{2}(Y_l-Y_u)&=jY_y(\omega)\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (317)$$

These two admittances are really the symmetrical and skew-symmetrical components of the asymmetric sideband characteristics, which were mentioned at the start of this section. It should be noted that $jY_y(\omega)$ is the admittance $Y_y(\omega)$ with its $-90°$ phase intercept removed; this phase intercept, at zero frequency, has been pointed out in equation 296 and illustrated by Figs. 102 (b) and 104 (b).

The equivalent modulation-frequency characteristics, $Y_x(\omega)$ and $Y_y(\omega)$, may therefore be obtained by the following process:

(*a*) Replot the asymmetric characteristics, as real and imaginary parts, in terms of the variations of the sidebands in relation to the carrier as a datum: i.e. in terms of $A_l/A_0$, $A_u/A_0$, $(\phi_0-\phi_l)$, $(\phi_u-\phi_0)$.

(b) Plot the mirror image of these characteristics, about the carrier frequency.
(c) Add the real parts of these two characteristics and then add the imaginary parts, giving $2Y_x(\omega)$.
(d) Subtract the real parts and then subtract the imaginary parts, giving $2jY_y(\omega)$.

This process is illustrated by Fig. 109, showing the symmetrical and skew-symmetrical components. These components have been converted from their complex form back into their modulus and phase-shift forms; the complex form must be used for the addition and subtraction operations. It should be noted that we have not transferred the carrier to zero in this process and that, strictly speaking, these components are the *band-pass equivalents* of $Y_x(\omega)$ and $Y_y(\omega)$ obtained by transferring these up from zero frequency (see Fig. 102 (b)) to the carrier frequency $\omega_0$.

The reader will appreciate that this process of adding or subtracting an asymmetrical characteristic and its own mirror image is identical with the process described in Chapter 2, Sec. 20, for obtaining the sine and cosine (or odd and even) components of a complex spectrum. There is really no essential difference between a spectrum and the frequency characteristics of a channel: both are curves, plotted on a frequency scale, of the amplitudes and phase shifts of a range of sinusoidal components, usually continuous.

## 67. Envelopes with rapid build-up rates

Many communication channels are in use to-day in which extremely short envelope build-up times are involved; such channels are capable of giving signal responses with extremely good "definition." It can happen, for example due to frequency changing, that the envelope is required to build up or decay at a rate which is comparable to the rate of change of the carrier wave itself. That is to say, the time of build-up or decay is comparable with one quarter period of the carrier.

In such a modulated wave the envelope spectrum contains frequencies which are so high that they approach the carrier frequency. Looked at another way, the spectrum of the modulated wave must contain sidebands which extend nearly to zero frequency and nearly up to twice carrier frequency. The band-pass filters designed to transmit such signals must have band-widths comparable to the mid-band frequency. Alternatively it is sometimes

the practice to use wide-band low-pass filters for transmission of such signals.

The question then comes up: what is the exact meaning to be attached to the word "envelope" in such cases, and what is the envelope waveform which results when such a modulated carrier wave is transferred to zero frequency? Also, when calculations have been made of the asymmetric sideband distortion of such signals, can the calculated envelope be taken to represent the actual channel response?

It is easily shown that the "envelope" of such a signal, which builds up or decays in a time comparable with a quarter carrier period, must be taken to mean the envelope in the true mathematical sense. This is *not* the locus of the peaks of the modulated wave, but is the curve which touches the modulated wave and has the same slope as this wave at every point of contact, independent of the carrier phase. If this meaning be attached to the term "envelope," whenever it has been used in this Chapter or elsewhere, then the theorem concerning the transfer of the carrier to zero frequency and the development of the "equivalent modulation-frequency characteristics" may be put to practical use, however great the build-up rate of the signal may be.

Let us now prove this point. We have written in equation 290 the expression for a carrier wave modulated by a signal function $F(\omega t)$ as:

$$e_1 = E \,.\, F(\omega t) \,.\, \cos \omega_0 t \quad . \quad . \quad . \quad . \quad [(290)]$$

and it is this wave, $E \,.\, F(\omega t)$, which has been termed the "envelope" and the distortion of which we have investigated. [Such a modulated wave may be produced in practice by applying the modulating signal $F(\omega t)$ to a balanced modulator with a square-law modulation characteristic.[4]]

At the points common to both waves, the envelope $E \,.\, F(\omega t)$ and the modulated wave $E \,.\, F(\omega t) \,.\, \cos(\omega_0 t + \phi)$ (where $\phi$ is any phase angle), we have:

$$E \,.\, F(\omega t) = E \,.\, F(\omega t) \,.\, \cos(\omega_0 t + \phi)$$

giving

$$t = (N\pi - \phi)/\omega_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (318)$$

where $N = 0, 2, 4, 6 \ldots$ any even number.

Also the slope of the modulated wave is:

$$\frac{de}{dt} = -\omega_0 E \,.\, F(\omega t) \sin(\omega_0 t + \phi) + E \,.\, \frac{dF(\omega t)}{dt} \,.\, \cos(\omega_0 t + \phi)$$

which at the common points, given by 318 above, becomes:

$$\frac{de}{dt}=E \cdot \frac{d \cdot F(\omega t)}{dt} \quad . \quad . \quad . \quad . \quad . \quad . \quad (319)$$

Thus the envelope is the curve which touches the modulated carrier and has the same slope at the points of contact, whatever the carrier phase angle $\phi$.

This implies, of course, that the modulated wave must be considerably distorted from the pure sinusoidal shape in the regions where it is building up or decaying rapidly. Fig. 110 illustrates a

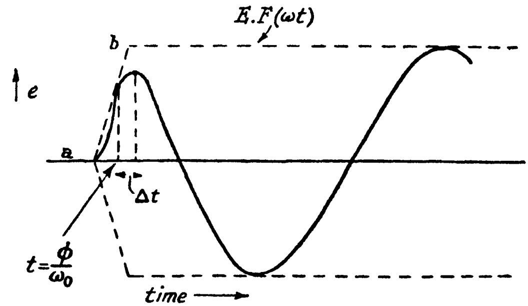



Fig. 110.—Modulated Wave with Envelope Rising at a Rate Greater than that of the Unmodulated Carrier.

modulated wave which has a rate of build-up even greater than that of the unmodulated carrier itself, showing the distortion that the wave undergoes in order to fit under the envelope, which is given by the dotted line. It is seen that during the time of build-up the wave touches the envelope $E \cdot F(\omega t)$ at a point $\Delta t$ before the peak of the carrier cycle; this time $\Delta t$ increases with the rate of build-up, but is a negligible quantity for normal modulation, in which the build-up time occupies many carrier cycles and the envelope becomes nearly identical to the locus of the wave peaks.

## REFERENCES

1. W. J. Poch and D. W. Epstein, "Partial Suppression of One Sideband in Television Reception," *Proc. I.R.E.*, Vol. 25, p. 15, 1937.
2. S. Goldman, "Television Detail and Selective-Sideband Transmission," *Proc. I.R.E.*, Vol. 27, p. 725, 1939.
3. L. S. Nergaard, "A Theoretical Analysis of Single-Sideband Operation of Television Transmitters," *Proc. I.R.E.*, Vol. 27, p. 666, 1939.

4. E. C. Cherry, "The Transmission Characteristics of Asymmetric Sideband Communication Networks" (Part 1), *J.I.E.E.*, Vol. 89, Part III, p. 19, 1942; and (Part 2), Vol. 90, Part III, p. 75, June 1943.
5. P. P. Eckersley, "Asymmetric Sideband Broadcasting," *J.I.E.E.*, Vol. 83, p. 36, July 1938; also Vol. 77, p. 517, 1935; also *Proc. I.R.E.*, Vol. 26, p. 1041, Sept. 1938.
6. N. Koomans, "Asymmetric Sideband Broadcasting," *Proc. I.R.E.*, Vol. 27, p. 687, Nov. 1939.
7. R. Urtel, "Bemerkungen zum Einseitenbandbetrieb im Fernsehen," Telefunken, July 1939.
8. J. M. Hollywood, "Single Sideband Filter Theory with Television Applications," *Proc. I.R.E.*, Vol. 27, p. 457, July 1939.
9. H. Nyquist and K. W. Pfleger, "Effect of the Quadrature Component in Single Sideband Transmission," *Bell Syst. Tech. Jour.*, Vol. 19, p. 63, 1940.
10. R. D. Kell and G. L. Fredendall, "Selective Sideband Transmission in Television," *R.C.A. Review*, Vol. 4, p. 425, 1940.
11. F. M. Colebrook, "The Frequency Analysis of the Heterodyne Envelope," *Exp. Wireless* (*Wireless Engineer*), Vol. 9, p. 195, April 1932.
12. F. E. Terman, "Fundamentals of Radio," Chap. 3, McGraw-Hill, 1938.
13. J. E. Smith, B. Trevor, and P. S. Carter, "Selective Sideband versus Double Sideband Transmission of Telegraph and Facimile Signals," *R.C.A. Review*, Vol. 3, p. 213, 1938.
14. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans. A.I.E.E.*, Vol. 47, p. 617, 1928.
15. C. C. Eaglesfield, "Carrier Frequency Amplifiers—Transient Response with Detuned Carrier," *Wireless Engineer*, Vol. 23, No. 270, Mar. 1946.
16. D. G. Tucker, "Transient Response of Filters," *Wireless Engineer*, Vol. 23, p. 36, Feb. 1946, and p. 84, Mar. 1946.

# CHAPTER 8

# Reflection and Echo Effects in Lines and in Lumped Networks

## 68. Examples of echo production

The phenomenon of reflection in a mis-matched line or cable and the appearance of echoes in these and in artificial lines and delay networks are familiarities nowadays. Strictly speaking, echoes can only be produced in extended systems, that is, in systems such as transmission lines, which have *length* and along which energy can be sent only at a finite velocity; this applies not only to electrical systems but to sound and other waves also. However, a study of the behaviour of transmission lines and of the way in which reflection and echo effects arise can lead to a better understanding of the characteristics of delay networks, filters, and other networks, which need have no significant length but which may be considered as *lumped* circuits. Furthermore, there is an interesting method of transient response analysis based on the idea of echoes in electric circuits, which will be described in this chapter, the application of which is not limited to extended systems but refers to any lumped circuit. This method is based on the steady-state frequency characteristics, and although it has its limitations as an exact method of analysis it provides another example of a means of obtaining an approximate result, though the approximation may be made as close to the exact solution as the user desires. Taken at its very least value, as a method of analysis, the idea of electric echoes can provide a means of estimating transient distortion from an inspection of the frequency characteristics of a network, at any rate as regards the general type and degree of the distortion.

Reflection does not often arise in power transmission* systems, in which the frequency is very low, 50 or 60 cycles/sec., unless the path of transmission is extremely long, but it does appear and can cause trouble in telephone lines, working at speech frequencies. It was observed, as soon as long telephone lines came into use, that serious signal distortion could be set up due to the reflected energy

* See, for example, H. Rissik, "Power System Interconnections," Chapter VI, Pitman & Sons.

arriving back and interfering with the sent signal, with a delay time of perhaps a whole syllable. Means are now usually provided for checking the reflected energy, in long telephone lines, by means of "echo suppressor" circuits. More recently, effects have been observed in television channels due to echoes being set up in the aerial feeder system,[1] partly by virtue of irregularities in the feeder construction and partly because the aerial impedance does not provide a correct match for the feeder at all transmitted frequencies. This effect can give rise to a second picture on the television receiver screen, superimposed on the wanted picture but displaced slightly in the direction of scanning. The same trouble sometimes occurs in picture telegraphy, in which the picture signals being transmitted along a cable are reflected by a mis-matched impedance at the receiver end, sent back to the transmitter and reflected again by a mis-match back to the receiver.*

Those engineers who have been concerned with short- or micro-wave transmission will be familiar with the reflections which arise in mis-matched feeders or waveguides. In such cases of steady-state transmission the reflected wave from the mis-matched end interferes with the forward wave, setting up a series of maxima and minima of voltage and current along the line, termed a stationary wave; the modulus and phase angle of the mis-matched load (possibly an aerial) determine the ratio of any one maximum to a minimum, called the standing-wave ratio, together with their positions relative to one end of the line.[3, 4]

Turning now to lumped networks, the artificial line or delay network has been in common use during recent years in radar circuits. Here the phenomenon of reflection has been put to practical use for the generation of short pulses. Lengths of transmission line or cable have also been used for the same purpose. Fig. 111 illustrates the application: a step wave is applied to the input end of a delay network or cable, possibly by the sudden connection of a charged condenser, and this step wave travels along the line and is reflected from the short-circuited end back again to the input terminals. The short-circuit provides an extreme case of a mis-match and *all* the energy is reflected back, but in this case the reflected step wave is reversed in sign by the short-circuit. The negative reflected step wave arrives back at the input terminals after a certain interval of time $\Delta t$, which depends on the length of

* A photograph of a received picture in such a case is shown on p. 156 of reference 2, given at the end of this chapter.

the cable or on the effective length of the delay line. This reflected step-wave subtracts from the applied step wave, thus forming a rectangular pulse of duration $\Delta t$—at least in theory, though there are many difficulties in the practical design of such circuits.

Similarly an open-circuit acting as the reflecting barrier would send back an echo of the step wave, but in this instance of the *same* sign. This setting up of positive or negative echoes is just given as a fact here, but is examined later in Sec. 71.

Now we have referred here to echoes in both a transmission line, which is an extended circuit and in which energy can move only

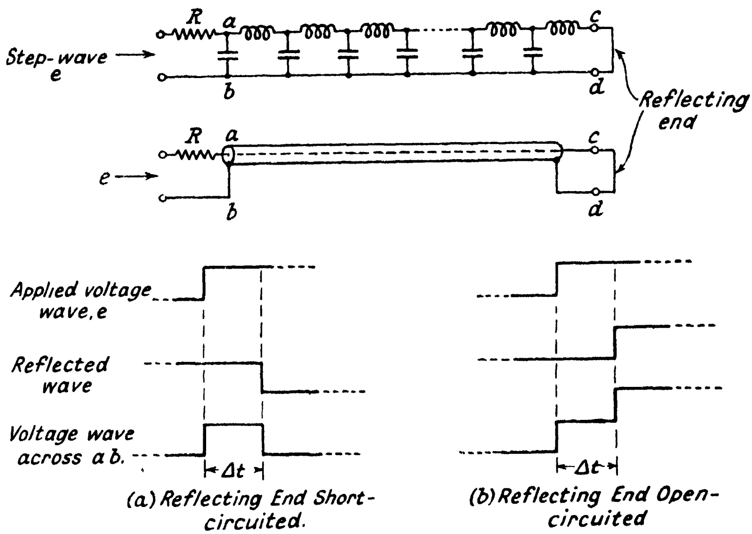



Fig. 111.—Reflections of Step Wave in a Transmission Line or a Delay Network.

with a finite velocity, and in a delay network (illustrated by a constant-$K$ low-pass filter structure in Fig. 111), which is a lumped circuit and in which the idea of a velocity is meaningless, there being no length necessarily involved. In the case of such lumped circuits any apparent time taken by a signal to travel through it can arise only by virtue of the phase-shift characteristic, as an envelope delay. This effect has received attention in Sec. 40 and need not be examined further here. Any lumped circuit, such as a delay network, must give some response at its output terminals instantaneously with the application of a signal to its input terminals; however, as often happens, this response can be very small until a certain length of

time has elapsed after which the signal builds up to its full value. This length of time, corresponding to $\Delta t$ in Fig. 111, may be called a *virtual delay time* and takes the place of the definite rate of travel of a signal along a transmission line; similarly, the echoes produced by such lumped circuits may be termed *virtual echoes* to distinguish them from the real thing.

The relations between transmission lines and their artificial lumped equivalents, together with the similarity of their behaviour in producing echo effects, suggests that it may be beneficial to extend the theory of echoes not only to artificial lines and delay networks but to filters and other structures which are derived from the artificial line, as a method of examining their transient response characteristics.

Before proceeding, it will be well to inspect a little more closely the relations between extended and lumped circuits in order to check the foundations on which this method is based.

## 69. Lumped and extended circuits

The elementary theory that has been developed for steady-state propagation along a transmission line is based on certain assumptions, which are justified by practical results.[5] The most important of these assumptions are geometrical and electrical uniformity along the whole length of line, and primary constants which are linear and independent of frequency; these primary constants are series inductance and resistance of the lines and shunt capacity and leakage between the lines, measured over a unit length. The equations relating the current and voltage at any point in the line are then developed by consideration of a short length $\delta x$ of the line, at a distance $x$ from the sending end.

Such a short length of line may be represented by a lumped 4-terminal circuit of the form shown in Fig. 112 (a), at one particular frequency only. The element values are found by measuring the series inductance and resistance and the shunt capacity and leakage of the piece of line, length $\delta x$, at a given frequency; if such a measurement is made, the element values will be found to vary with frequency owing to the finite length of this piece of line. This circuit is then similar in behaviour to the length of line at one particular frequency only, and is not a complete equivalent circuit (see Sec. 28).

Let these element values, as measured, be $L_1$, $R_1$, $C_1$, and $G_1$. Since they have been measured with a length of line $\delta x$, we may

divide them by $\delta x$ to obtain what may be called the lumped equivalent inductance, resistance, capacity, and leakage per unit length $L$, $R$, $C$, and $G$, for the particular frequency, so that the element values in the actual circuit are as shown in Fig. 112 (a).

By Kirchhoff's laws we may write down the equations relating voltage and current, in this circuit:

$$\left.\begin{aligned} -\delta v &= (R\,.\,\delta x)i + (L\,.\,\delta x)\frac{di}{dt} \\ -\delta i &= (G\,.\,\delta x)v + (C\,.\,\delta x)\frac{dv}{dt} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (320)$$

Or, putting in the actual measured values $L_1$, $R_1$, $C_1$, and $G_1$ instead:

$$\left.\begin{aligned} -\delta v &= R_1 i + L_1\frac{di}{dt} \\ -\delta i &= G_1 v + C_1\frac{dv}{dt} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (321)$$

These equations represent the behaviour of the lumped circuit, Fig. 112 (a), whatever the waveforms of $v$ and $i$, but of the length

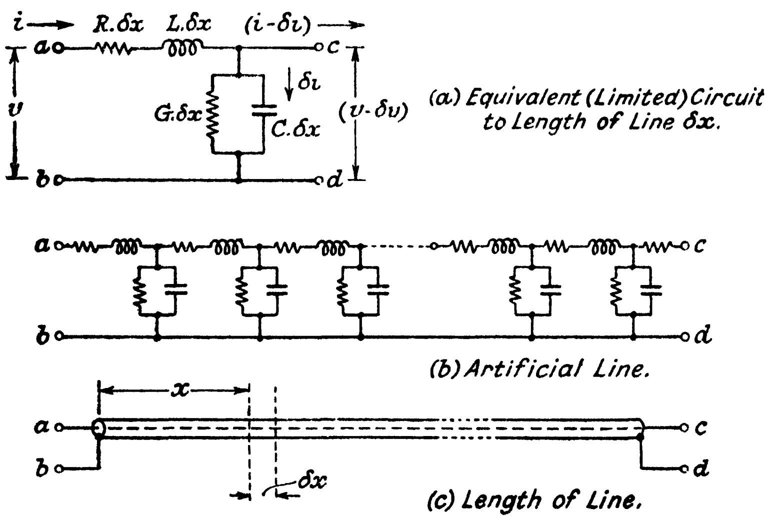



Fig. 112.—Length of Line and its Lumped Equivalent.

of line only for the steady-state, at one particular frequency. The transient responses of the short length of line and of the lumped equivalent circuit are therefore not identical.

The complete line may be represented by a number of such equivalent circuit sections connected in cascade, Fig. 112 (b); the

number required depends on the lengths, $\delta x$, of short pieces into which the line has been considered to be divided, each length $\delta x$ being replaced by a single section of the lumped equivalent circuit. Such a structure, or modifications of it,* is called an *artificial line* [11] and has certain properties similar to those of the real line, which have been mentioned in the preceding section.

This artificial line simulates the real line (i.e. it has identical input, output, and transfer impedances) at one particular frequency only, but nevertheless it approximates to the real line over a certain band of frequencies.[2] If the length $\delta x$, simulated by one section, is now made infinitesimal, then the number of sections in the artificial line increases towards infinity. The greater the number of sections, simulating a given length of line, the greater the band of frequencies over which the properties of the artificial and the real line correspond and the closer their transient responses become. It should be noted that when $\delta x \to 0$ the number of sections becomes infinite; it is then *identical* with the real line.

Equations 320 may then be modified so that they give the current and voltage, $v$ and $i$, at any point $x$ in the length of line. Dividing both sides of the equations by $\delta x$ and letting $\delta x \to 0$ gives:

$$\left.\begin{aligned} -\frac{\partial v}{\partial x} &= Ri + L\frac{\partial i}{\partial t} \\ -\frac{\partial i}{\partial x} &= Gv + C\frac{\partial v}{\partial t} \end{aligned}\right\} \quad . \; . \; . \; . \; . \quad (322)$$

These equations are now true for the length of line, whatever the waveforms of $v$ or $i$.

The mathematical reason why real echoes cannot appear in artificial lines, but only in real lines, is now apparent.† The above equations, 322, involve the length variable $x$, whilst equations 321, representing the behaviour of the artificial line section, do not. This is the essential difference between these equations representing the currents and voltages in the two systems—the extended and the lumped circuits.

* For example, if series resistance $R$ and shunt leakage $G$ are negligible, this artificial line become the constant-$K$ delay network of Fig. 111.

† This distinction between real echoes in transmission lines and virtual echoes in artificial lines may seem an academic point here. It is of primary importance when we come to examine the transient response of non-periodic circuits in terms of the echo idea.

The behaviour of artificial lines and other structures of similar type, such as wave-filters, is closely allied to that of the transmission line, as would be expected from the above elementary derivation of the "equivalent" circuit to a line. Many filter structures bear little resemblance to our artificial line in their circuit configurations, but they are nevertheless derivable, in their basic forms, from the transmission line and its simulation by lumped circuits. For example the structure in Fig. 112 (b) becomes a constant-$K$, low-pass filter if the resistance elements be removed; such a structure, with its filtering properties, was originally discovered as a result of inserting loading-coil inductances in telephone lines, thereby producing a circuit partly lumped and partly extended.

As has already been stated, artificial lines are equivalent to a length of transmission line at one frequency only, though they approximate over a certain band of frequencies. Wave-filters on the other hand, which ideally possess no dissipative elements, are equivalent to a transmission line over a band of frequencies which may be made as wide as required; outside that band the resemblance ceases and the filter exhibits its cut-off effect, this being an essential difference between a transmission line and its lumped equivalent.

If the transmission line is absolutely perfect and produces no signal distortion its frequency characteristics must be uniform—the modulus must be a constant at all frequencies and the phase-shift characteristic must be of a constant slope (see Sec. 38). Ideally a wave-filter has similar characteristics up as far as the cut-off frequency (as illustrated by full lines in Fig. 113), but, as we have already seen from our consideration of idealised filters in Chapter 5, such characteristics would imply a response signal having a finite amplitude before $t=0$, the time of application of the input signal, which is clearly an impossible condition. In practice, filter characteristics will be found to depart slightly from these idealised forms, even over the pass band, and to have a slight waviness,[7] like that shown in the same figure, Fig. 113,* by the dotted lines.

This departure from the idealised form, essential in any practical filter, must be such as to remove all traces of signal before $t=0$, but also sets up distortion of the response signal. This distortion may be conveniently considered to be due to "virtual" reflections from the filter termination, caused by an imperfect match; any filter with a finite number of sections must be mis-matched by its termination over part of the pass-band. The only way to terminate

* For example see the practical filter characteristics in Figs. 33 and 57.

perfectly is to use an infinite number of filter sections, a practical impossibility (the so-called *iterative termination*).

Thus the appearance of echoes in a filter transient response is bound up with this waviness or departure from uniformity on the

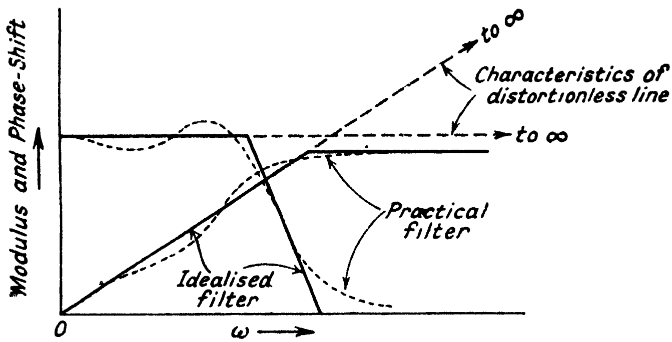



Fig. 113.—Frequency Characteristics of (a) Distortionless Line, (b) Idealised Filter, Uniform up to Cut-off, (c) Practical Filter, with Inherent Waviness.

frequency characteristics,* a relation which we shall look into more closely later and which we may extend to other kinds of lumped circuit.

## 70. Characteristics of transmission lines—reflectionless case

In this and the following sections we shall run through briefly that part of the simple steady-state transmission-line theory that is relevant to this study of reflection and echo effects. Transmission line theory is dealt with in great detail in many standard books and need not be reproduced at length here.[5, 6, 8]

As we have shown in Sec. 69, equations 322 represent the current and voltage at any point, distant $x$ from the sending end, in a length of line. These equations, though true in this form for transient as well as steady-state currents and voltages, simplify if merely the steady state is of interest. If $i$ and $v$ are purely sinusoidal, these equations may be rewritten:

$$\left.\begin{aligned} -\frac{dv}{dx} &= (R+j\omega L)i \\ -\frac{di}{dx} &= (G+j\omega C)v \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (323)$$

* See p. 498 of reference 8.

Differentiating both sides with respect to $x$ gives:

$$\left.\begin{aligned}\frac{d^2v}{dx^2}&=(R+j\omega L)(G+j\omega C)v\\ \frac{d^2i}{dx^2}&=(R+j\omega L)(G+j\omega C)i\end{aligned}\right\} \quad . \; . \; . \quad (324)$$

The solutions* of these equations may be shown to be of the form:

$$\left.\begin{aligned}v&=A_1\epsilon^{\gamma x}+A_2\epsilon^{-\gamma x}\\ i&=A_3\epsilon^{\gamma x}+A_4\epsilon^{-\gamma x}\end{aligned}\right\} \quad . \; . \; . \; . \; . \; . \quad (325a)$$

where $\gamma$ is written for

$$\sqrt{[(R+j\omega L)(G+j\omega C)]} \quad . \; . \; . \; . \quad (325b)$$

and where the $A$'s are arbitrary constants which depend on the conditions at the end of the line. Let us assume that, at the *sending* end of the line (where $x=0$) $v=v_s$, and $i=i_s$. This gives two relations, but there are altogether four arbitrary constants. The other two are found by differentiating both sides of the equations 325a with respect to $x$ and substituting in the original equations 323:

$$\left.\begin{aligned}\gamma\{A_1\epsilon^{\gamma x}-A_2\epsilon^{-\gamma x}\}&=-(R+j\omega L)i\\ \gamma\{A_3\epsilon^{\gamma x}-A_4\epsilon^{-\gamma x}\}&=-(G+j\omega C)v\end{aligned}\right\} \quad . \; . \quad (326)$$

inserting the conditions $v=v_s$ and $i=i_s$ at $x=0$ then gives two other relations from which the four arbitrary constants may be found. In terms of these sending-end conditions the voltage and current, at any point $x$, then appear as:

$$\left.\begin{aligned}v&=v_s\cosh\gamma x-i_sZ_0\sinh\gamma x\\ i&=i_s\cosh\gamma x-\frac{v_s}{Z_0}\sinh\gamma x\end{aligned}\right\} \quad . \; . \; . \quad (327)$$

where $Z_0$ is written for:

$$\sqrt{\left(\frac{R+j\omega L}{G+j\omega C}\right)} \quad . \; . \; . \; . \; . \; . \; . \quad (328)$$

$Z_0$ has the dimensions of an impedance, and it may easily be shown that it is the impedance looking into the sending end (i.e. across the sending terminals) of an infinite length of the line. It should be noted that the equations (327) are quite general and refer to any length of line with any termination at the receiving end.

* These are simple second-order linear differential equations, such as we have examined in Chapter 1, the solutions being exponential in form.

Let us take this special case of an infinitely long line. Such a line must have infinite attenuation at the hypothetical point $x=\infty$, however small the loss elements $R$ and $G$ are, so that $v=0$ and $i=0$ there. Inserting these conditions in either of the equations 327 gives:

$$\frac{v_s}{i_s}=Z_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (329)$$

and $v_s/i_s$ is the sending-end impedance, so that we have established our point concerning $Z_0$. Furthermore, $Z_0$ will be the sending-end impedance of any finite length of line L if this be terminated by an impedance equal to $Z_0$; any such length L may be cut off an infinitely long line, and the remainder will still be infinitely long and have a sending-end impedance of $Z_0$ and hence this remainder may be simulated by a lumped impedance, equal to $Z_0$, connected across the receiving-end terminals of the length L. For this reason $Z_0$ is called

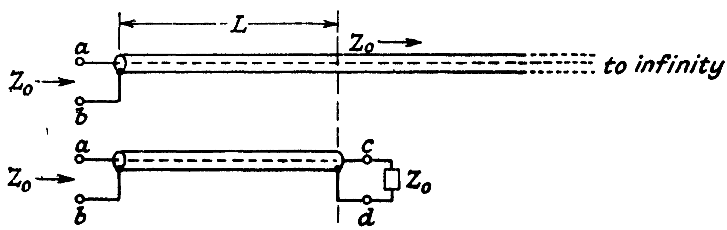



Fig. 114.—A Length of Line Terminated in its Characteristic Impedance $Z_0$.

the *characteristic impedance* of the line. Fig. 114 illustrates this point.

Such an impedance, $Z_0$, is said to *match* the line of characteristic impedance $Z_0$, and a length of line L so terminated behaves, over that length, as though it were continued to infinity. For example, a signal may be transmitted along such a line and all the energy, as it reaches the end, is absorbed in the termination. If we consider the length of line to be a generator of internal impedance $Z_0$ driving a load (the terminating impedance) then the maximum amount of energy is transferred to this load when its impedance is made the same as the generator impedance $Z_0$. Any other load absorbs less energy and the reduction may be considered to be caused by a reflection from this mis-matching load, sending energy back into the generator (the cable).

If a length of line is correctly matched no reflection of energy can occur. In these circumstances the equations 327 simplify and the

voltage and current at any point in an infinitely long line, or correctly matched line, may be expressed as:

$$\left.\begin{aligned} v &= v_s\epsilon^{-\gamma x} \\ i &= i_s\epsilon^{-\gamma x} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (330a)$$

(using 329). Both $v$ and $i$ have the same form in this matched case and their ratio $v/i$ is constant at every point in the line and equal to $Z_0$, since from (330a):

$$\frac{v}{i}=\frac{v_s}{i_s}=Z_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (330b)$$

Now the term $\gamma$, usually called the *propagation constant*, is of special significance.* From (325b) we see it is a complex quantity and its real and imaginary parts may be separated:

$$\begin{aligned}\gamma=&\sqrt{\tfrac{1}{2}}\{\sqrt{[(R^2+\omega^2L^2)(G^2+\omega^2C^2)]}+(GR-\omega^2LC)\} \\ &+j\sqrt{\tfrac{1}{2}}\{\sqrt{[(R^2+\omega^2L^2)(G^2+\omega^2C^2)]}-(GR-\omega^2LC)\}. \quad (331)\end{aligned}$$

which may be written:

$$\gamma=A+jB \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (332)$$

These expressions, $A$ and $B$, are ungainly, but notice that they differ only in a sign.

In terms of these real and imaginary parts of $\gamma$ equations 330 become:

$$\left.\begin{aligned} v &= v_s\epsilon^{-Ax} \ . \ [\cos Bx+j\sin Bx] \\ i &= i_s\epsilon^{-Ax} \ . \ [\cos Bx+j\sin Bx] \end{aligned}\right\} \quad . \quad . \quad (333)$$

From these the vectors may be drawn representing the voltage $v$ and current $i$ at any point in the line. It can be seen that both $v$ and $i$ (the two expressions are identical apart from magnitude) decay exponentially with distance along the line, $x$, and that the phase shift is proportional to $x$. The position and length of the vector, for $v$ or $i$, at any point $x$ is thus given by a vector diagram of the type illustrated in Fig. 115. The expression $A$ determines the rate at which $|v|$ or $|i|$ decay with distance, and is termed the *attenuation constant*,* while $B$ determines the phase shift per unit distance and is called the *phase constant* or sometimes the *wavelength constant*.* The vector rotates $2\pi$ over a length of line $x=2\pi/B$, which may be termed a *wavelength*.

Fig. 115 (a) shows the "end elevation" of this three-dimensional vector diagram, a logarithmic spiral, with three typical vectors at

* It is *constant* at any point in the line, as regards $x$, but of course it is a function of frequency.

the points $x=0$, $x_1$ and $x_2$. The line joining the ends of any two vectors, say $x_1$ and $x_2$, is itself a vector representing the transverse voltage measured between these two points, $x_1$ and $x_2$, on the line.

Similarly we could construct a vector diagram for the artificial line in Fig. 112 (b), which would, however, consist of a series of discrete vectors, one at the end of each section, lying on the same type of spiral locus. Fig. 115 (c) illustrates the vectors for the first two sections of such a network. Comparison of the diagrams for the real and the artificial lines shows that *distance* $x$ in the first corresponds to *section number* in the second; a section of the

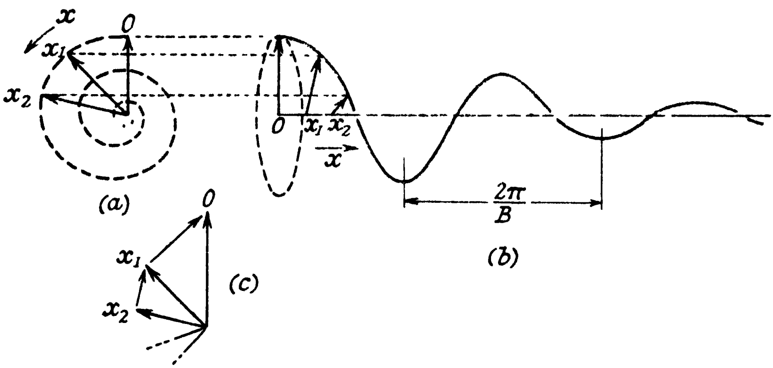



Fig. 115.—Vector Diagram for Lossy Transmission Line.
(a) End Elevation. (b) Locus along Line.
(c) End Elevation for Artificial Line, Two Sections.

artificial line was derived, originally, from the measured constants of a length $\delta x$ of line at a particular frequency $\omega$.

This vector diagram is a *stationary* one and represents the voltage and current distribution along the line at a given instant of time. If this vector diagram is now considered to rotate at a uniform angular velocity, $\omega$, the variations of current and voltage with time will be evident. The spiral locus, Fig. 115 (b), will move with a wavelike motion; the current and voltage at any point $x$ will vary sinusoidally (frequency $\omega$); and the peak amplitude will be given by the vector length at that point. During one period, $2\pi/\omega$, the wave travels one wavelength, $2\pi/B$, so that the velocity of the wave along the line is $\omega/B$ cm./sec. (assuming that $B$ is measured in radians/cm., or that the primary constants $R$, $L$, $G$, and $C$ are specified per cm. length of line).

This velocity is the *steady-state wave velocity*, sometimes called *phase velocity*, and it is not the velocity with which a pulse, a transient, or other signal, may be transmitted that may constitute information or intelligence. A pure sinusoidal wave cannot convey information, as has been emphasised already. The velocity with which a pulse or transient signal travels is called the *group velocity*, and is in general a lower one.

This distinction corresponds to that between phase delay and envelope (or group) delay, with which we dealt in Sec. 40 and illustrated by Fig. 62. Suppose the curve in Fig. 62 (a) represents the phase-shift/frequency characteristic of our line. Then the phase shift per unit length of line (which we have called $B$, measured at one particular frequency $\omega$) is a function of $\omega$, as represented by this characteristic, and may be written $B(\omega)$. The steady-state wave velocity, at any one frequency $\omega$, is then $\omega/B(\omega)$, but the velocity with which the signal corresponding to a narrow spectrum, $\omega \pm \delta\omega$, is transmitted is $\delta\omega/\delta B(\omega)$, the reciprocal of the slope of the tangent to the phase-shift curve at the point $\omega$. This signal is a modulated wave, carrier frequency $\omega$, of pulse-like form having the envelope shown in Fig. 62 (b).

The group velocity $\delta\omega/\delta B(\omega)$ is clearly the reciprocal of the group (or envelope) delay time $\delta B(\omega)/\delta\omega$; it is a *velocity* since the phase-shift characteristic, $B(\omega)$, is measured over a unit length of line. If this characteristic $B(\omega)$ is measured for one section of an artificial line then $\delta B(\omega)/\delta\omega$ is the envelope delay time through that one section, and the reciprocal $\delta\omega/\delta B(\omega)$ is the virtual group velocity through the artificial line, in sections per second.

This signal, corresponding to a narrow uniform spectrum extending over the frequency range $(\omega \pm \delta\omega)$, has been illustrated by Fig. 62 (b) as a function of time, and if such a signal be applied to the sending-end terminals of a line, it will only be propagated along the line with the velocity $\delta\omega/\delta B(\omega)$ and will appear at the receiving-end terminals. However, we may imagine this signal to be distributed spacially along the line, during its travel, as a function of $x$. In the case of the artificial line this idea is, of course, meaningless.

## 71. Characteristics of transmission lines—showing reflection

In the previous section we have dealt with transmission along a line of quite general characteristics, $R$, $L$, $G$, and $C$ having any values. In practice a transient signal travelling along a length of line is distorted in shape owing to the fact that the propagation "constant"

$\gamma$ varies with frequency (equation 331). The group (or envelope) delay $\delta B(\omega)/\delta\omega$ thus varies with frequency, and different parts of the signal spectrum are transmitted with different velocities.

There are certain special cases that are of interest:

(a) *The Distortionless Line.*

If it were possible to relate the primary constants* so that the time-constants of the series arms ($R/L$) and shunt arms ($G/C$) are equal (see Fig. 112):

$$\frac{R}{L}=\frac{G}{C} \quad \dots\dots\dots \quad (334)$$

then substitution of this relation in equation 331 gives for the propagation constant:

$$\gamma=A+jB=\sqrt{(RG)}+j\omega\sqrt{(LC)} \quad \dots \quad (335)$$

In this case the attenuation $A$ is independent of frequency and the phase constant $B$ is linear and proportional to frequency. These are the idealised distortionless conditions (Sec. 38), illustrated by Fig. 113 (the dashed curves), and a signal of any kind would be transmitted along such a line without waveform distortion and with a velocity $\delta\omega/\delta B = 1/\sqrt{LC}$ a constant. Note the phase and group velocities are identical in this case.

The value of $Z_0$, the characteristic impedance, becomes a pure resistance. By inserting the condition 334 for distortionless transmission, equation 328 may be written:

$$Z_0=\sqrt{\frac{L(G/C+j\omega)}{C(R/L+j\omega)}}=\sqrt{\frac{L}{C}} \text{ or } \sqrt{\frac{R}{G}} \quad \dots \quad (336)$$

which is usually written $R_0$.

Obviously these ideals cannot be attained in practice, though they can be approached and great improvement in transmission quality obtained by increasing $L$, since with most lines and cables $R/L \gg G/C$. The inductance can be increased either by aid of iron wire, tape, etc., in which case the primary "constants" vary with frequency, or by spiralling the conductor(s), in which case the capacity between adjacent turns introduces a new element and the original equations, 323, no longer hold. The usual practice is the insertion of inductance coils [9, 10] at intervals along the line length, which makes the whole system partly a distributed and partly a lumped one.[9] The effect of these lumped elements is to give uniform

* Assuming they themselves do not vary with frequency.

transmission characteristics up to a limiting frequency and then to introduce a cut-off.

(b) *The Lossless Line.*

If there is considered to be no dissipation in the line, $R=0$ and $G=0$. In this case the propagation constant becomes, from 331:

$$\gamma=A+jB=j\omega\sqrt{(LC)} \quad . \quad . \quad . \quad . \quad (337)$$

so that the attenuation constant is zero and the phase constant is again linear and proportional to frequency, as for the distortionless line. Also, again from 328:

$$Z_0=\sqrt{(L/C)}=R_0, \text{ a resistance} \quad . \quad . \quad . \quad (338)$$

Such an ideal line transmits a signal, undistorted and unattenuated, with a velocity $1/\sqrt{(LC)}$. The equivalent artificial line (see Fig. 112 (b)) becomes a constant-$K$ low-pass filter, the resistance elements being negligible.

This is the network commonly used for delaying pulses or other signals in television and radar systems, and usually called a "delay network," to be discussed further in Sec. 73.

The energy equation for the general line, with $R$, $L$, $G$, and $C$ all finite and unrelated, is derivable from the original equations, 322, for the current and voltage at any point in the line. Multiplying (on both sides) the first equation by $i$ and the second by $v$ gives:

$$\left.\begin{aligned} -i\frac{\partial v}{\partial x}&=Ri^2+L\frac{\partial i}{\partial t}\,.\,i \\ -v\frac{\partial i}{\partial x}&=Gv^2+C\frac{\partial v}{\partial t}\,.\,v \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (339)$$

adding these, and rewriting the differential coefficients:

$$-\frac{\partial}{\partial x}(vi)=(Ri^2+Gv^2)+\frac{\partial}{\partial t}(\tfrac{1}{2}Li^2+\tfrac{1}{2}Cv^2) \quad . \quad . \quad (340)$$

The term on the left-hand side represents the power being supplied in an infinitesimal length of the cable, while on the right-hand side we have the power lost in the resistance and leakage, $R$ and $G$, together with the rate of change of the energy (i.e. power) stored in the inductance and capacity, $L$ and $C$, in that infinitesimal length.*

* Compare this equation with the energy equation of a lumped circuit, such as equation 7. The element of *length* enters here.

These equations hold for any waveform of $v$ and $i$ and give the energy components of a pulse or other signal travelling along the line. If the line is of the distortionless or of the lossless type (see paras. (a) and (b) above) the ratio $v/i$ is constant at every point in the line and equal to $R_0$. But, from 336 and 338:

$$R_0 = v/i = \sqrt{(L/C)} \quad . \quad . \quad [(336) \text{ and } (338)]$$

$$\text{or } Li^2 = Cv^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (341)$$

so that in these ideal cases the energies stored in the inductance and in the capacitance elements are equal.

Consider the transmission of a short rectangular pulse signal along a length of line (see Fig. 116 (a)). If the pulse duration is $T_1$ it will "occupy" a length of line equal to $T_1/\sqrt{(LC)}$, the product

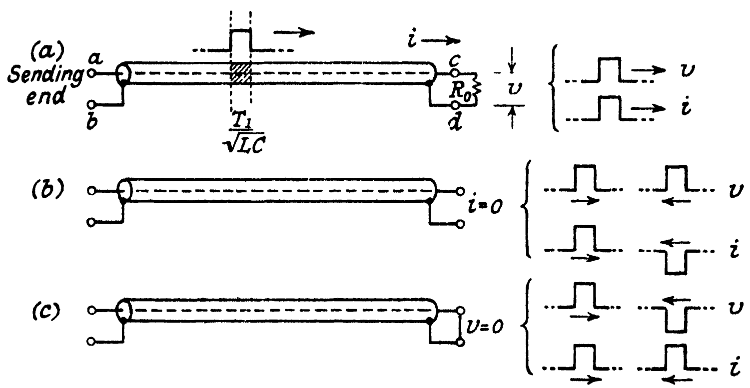



Fig. 116.—Pulse Echoes in a Line.

of time and velocity, which implies that energy is stored, at any instant, in the inductance and capacity over this length of line. If the line is a lossy one this energy is gradually reduced as the pulse travels along, but if the conditions 334 hold for distortionless transmission the pulse will preserve its rectangular waveform. If the line is finite in length, but terminated in its characteristic impedance $R_0$, as in (a), all the energy will be absorbed in $R_0$ and a current of rectangular waveform will flow through this resistance, setting up a voltage also rectangular in waveform. The pulse is then transmitted along the line without reflection.

If, however, as in Fig. 116 (b), the line is open-circuited at the receiving end, the pulse energy will be equally divided between $L$ and $C$ as it travels along the line, but when it reaches the open

circuit the magnetic energy must vanish, since there can be no current there. This energy cannot be lost, since there is no resistance to absorb it. It is therefore converted to electrostatic energy by a rise in the line voltage at this point. This rise in voltage is caused by the collapse of the magnetic field, which takes place at a *uniform rate* for a time equal to the pulse duration $T_1$, giving a constant induced voltage for this time interval.

The energy stored electrostatically at this instant cannot remain at the end of the line, since the line is a continuous conducting body, and so this energy travels back along the line as a reflected pulse, or echo. The sign of the reflected voltage pulse is thus of the same sign as the transmitted one, since the voltage is increased at the open-circuited end: i.e. the reflected pulse is momentarily added on to the arriving pulse, for a time equal to the pulse duration. The reflected *current pulse* must, however, be opposite in sign to the transmitted one since, at the open-circuit, the two component pulses must cancel, as there is zero magnetic energy there, again for a time equal to the pulse duration $T_1$.

By similar arguments the type of echo set up by a short-circuited end (Fig. 116 (c)) may be examined. There can be no electrostatic energy stored at this point and all the energy passes into the magnetic field, thereby increasing the pulse current in the short-circuit. The electrostatic energy again collapses at a uniform rate, taking a time $T_1$ to fall to zero. In this case the return voltage pulse is opposite in sign to the transmitted one, while the return current pulse is of the same sign, the two current pulses adding at the short-circuit for a time $T_1$.

For intermediate values of *resistive* mis-match, in between a complete open or short-circuit, some of the energy is absorbed in the termination and some is stored in *both* the electrostatic and magnetic fields. Depending upon whether the termination is greater or less than $R_0$, so the electrostatic or the magnetic energy, at that point, predominates and the returned echo sign is determined. The magnitude of the echo depends upon the amount of energy absorbed and is greater the greater the mis-match, either above or below $R_0$.

If the termination has reactance some of the energy of the pulse will be stored in this. In such cases an echo of the pulse is returned by the energy stored at the end of the line due to the mis-match, in the manner we have discussed, but also the energy held by the reactance of the termination will be released at a rate which depends on the type and configuration of the circuit elements forming the

termination. The returning pulse is thus deformed in shape and must certainly occupy a time longer than $T_1$, the exact waveform depending on the termination circuit structure.

There is an important distinction between the transmission of a pulse along the line and the steady-state transmission of a sinusoidal wave, particularly with the mis-matched line. We have already shown that the steady-state impedance at any point along a line is $R_0$, and is the ratio of the line voltage to current at the point if the line is *matched*. If it is not matched then the impedance departs from the value $R_0$ and in general varies from point to point along the line.[2, 3] However, when the pulse is travelling along the line it occupies only a certain length of the line, $T_1/\sqrt{LC}$ (see Fig. 116 (a)) and the line on either side is quite static and has no energy in it. This pulse forms a signal which is conveyed to the mis-matched end of the line, taking a finite time to do so, and hence it cannot be affected by the conditions at that end of the line until it reaches that point. The same is true of a step wave or any other transient signal. The pulse behaves as though it is travelling along a line of impedance $R_0$ until the instant it reaches the mis-matched end, whatever the termination impedance is at that end.

When a pulse, or any other signal, is reflected from the mis-matched end of a line, it travels back to the sending end. If this sending end is matched the returning pulse is absorbed completely; if not, a second echo is sent forward again, its magnitude and sign depending on the impedance across the sending end. If both ends are mis-matched a continuous series of echoes is set up, spaced apart by a time interval which depends on the length of the line and the velocity of propagation. Fig. 117 (d) shows the example of a line effectively open-circuited at both ends.

The case of particular interest is the line matched at the sending end and mis-matched at the receiving end—the single echo system. Fig. 117 (a) and (b) illustrate this; the line terminates in $R_0$ at the sending end and $R_T$ at the receiving end. If the generator supplies a pulse voltage of amplitude $E$, the voltage $v_s$ appearing across the actual sending-end terminals of the line is $E/2$, since the line itself appears to the pulse as a resistance $R_0$. The echo returns to the sending end after a time interval $t=2l\sqrt{(LC)}$, where $l$ is the line length and $1/\sqrt{(LC)}$ the velocity of propagation, i.e. a time corresponding to twice the length of the line. Fig. 117 (b) shows the voltage waveform of the forward pulse and echo for the two extreme cases $R_T=\infty$ and $R_T=0$.

## 72. Frequency characteristics of a reflecting line

The applied signal and its echo, shown in Fig. 117 (b), are measured at the sending-end terminals of the line. Let us now examine the sending-end impedance, as a function of frequency, measured at this point.

Fig. 117 (a) shows a length, $l$, of distortionless line terminated by

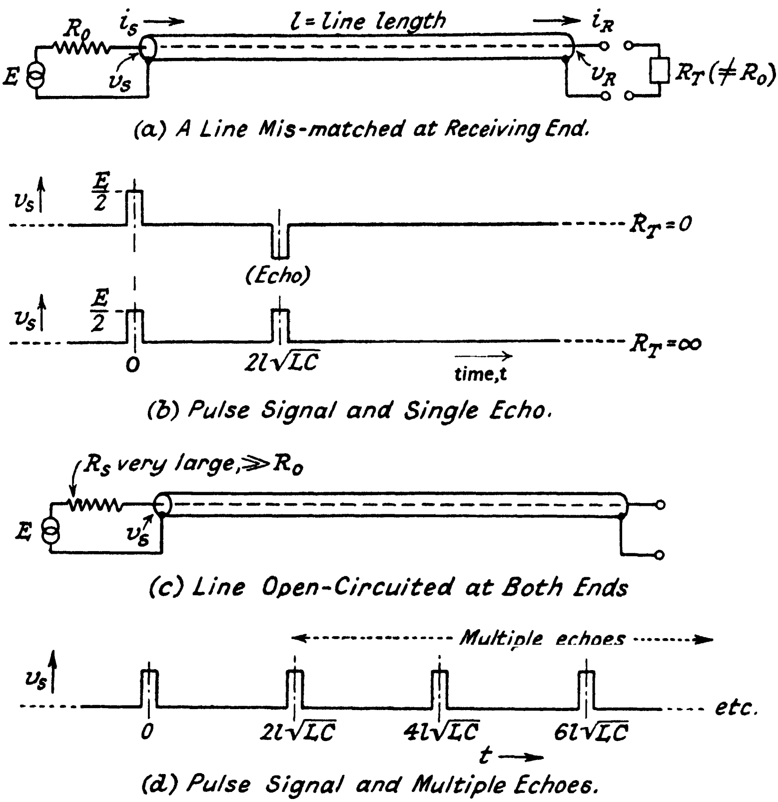



Fig. 117.—Line Showing Single and Multiple Echoes of a Pulse.

a resistance $R_T$, not necessarily equal to $R_0$. The voltage and current values at any point in a line are given by the general equations 327. At the termination, $x=l$, $v=v_R$, and $v_R=i_R R_T$. Inserting these conditions and eliminating $v_R$ gives:

$$\frac{v_s}{i_s}=Z_0\frac{R_T\cosh\gamma l+Z_0\sinh\gamma l}{Z_0\cosh\gamma l+R_T\sinh\gamma l} \quad . \quad . \quad . \quad (342)$$

But $v_s/i_s = Z_s(\omega)$, the sending-end impedance. We write this impedance as $Z_s(\omega)$ because we are specifically interested in it as a function of frequency.

Now if the line is distortionless, $Z_0 = R_0$, a pure resistance, and $\gamma = \sqrt{(RG)} + j\omega\sqrt{(LC)}$ from equations 335 and 336. Thus the hyperbolic functions, cosh and sinh, become circular ones:

$$\left.\begin{aligned}\cosh \gamma l &= \epsilon^{\sqrt{(RG)}} \,.\, \cos \omega l\sqrt{(LC)}\\ \sinh \gamma l &= j\epsilon^{\sqrt{(RG)}} \,.\, \sin \omega l\sqrt{(LC)}\end{aligned}\right\} \quad . \quad . \quad . \quad (343)$$

Substituting these conditions, that the line is distortionless, into 342 gives the sending-end impedance (the exponential terms vanish):

$$Z_s(\omega) = R_0 \frac{R_T \cos \omega l\sqrt{(LC)} + jR_0 \sin \omega l\sqrt{(LC)}}{R_0 \cos \omega l\sqrt{(LC)} + jR_T \sin \omega l\sqrt{(LC)}} \quad . \quad (344)$$

In the extreme mis-match cases, this modifies to:

(a) If $R_T = 0$, $\quad Z_s(\omega) = jR_0 \tan \omega l\sqrt{(LC)} \quad . \quad . \quad . \quad . \quad (345)$

(b) If $R_T = \infty$, $\quad Z_s(\omega) = -jR_0 \cot \omega l\sqrt{(LC)} \quad . \quad . \quad . \quad (346)$

Now this is the impedance looking into the line sending-end terminals, with no external circuit connected. The impedance 346 therefore corresponds to the line being effectively open-circuited at *both* ends, which sets up the voltage echo pattern illustrated by Fig. 117 (d), a continuous series of echoes. We shall be returning to this example later.

In the single echo case, Fig. 117 (a), the sending end is matched and has an impedance equal to $R_0$ across its terminals. Thus the complete impedance between these terminals consists of $R_0$ and $Z_s(\omega)$ in parallel; using the expression 344 for $Z_s(\omega)$, this impedance may be written:

$$\begin{aligned}\frac{R_0 \,.\, Z_s(\omega)}{R_0 + Z_s(\omega)} &= \frac{R_0}{2}\left[1 + \left(\frac{R_T - R_0}{R_T + R_0}\right) \cos 2\omega l\sqrt{(LC)}\right]\\ &\quad - j\frac{R_0}{2}\left[\left(\frac{R_T - R_0}{R_T + R_0}\right) \sin 2\omega l\sqrt{(LC)}\right] \quad . \quad . \quad (347)\end{aligned}$$

after rationalising.

This impedance is seen to vary with $\omega$ in a sinusoidal fashion, the amplitude of the variations depending on $(R_T - R_0)/(R_T + R_0)$ so that it increases with the mis-match, as $R_T$ and $R_0$ differ by an increasing amount. This sinusoidal shape of frequency characteristic is then associated with the production of a single echo; the steady-state impedance varies sinusoidally with frequency and the corresponding transient response consists of the applied signal with

a perfect echo. This type of system is unique and no finite electrical network, other than a length of line matched at the sending end and resistively mis-matched at the other end, has such characteristics. Although we have referred specifically to pulses and illustrated such signals in Fig. 117, any other type of signal is echoed without distortion by such a mis-matched line, the input-impedance/frequency characteristics of which are sinusoidal in shape.

Such perfect characteristics are unobtainable in practice,* [1] but are of great interest in analysis, as will be shown.

In the two extreme cases, $R_T=0$ and $\infty$, the sending-end impedance, 347, becomes:

(a) If $R_T=0$, impedance =

$$\frac{R_0}{2}\left[1-\cos 2\omega l\sqrt{(LC)}\right]+j\frac{R_0}{2}\left[\sin 2\omega l\sqrt{(LC)}\right] \quad . \quad (348)$$

corresponding to a single *negative* voltage echo, and

(b) If $R_T=\infty$, impedance =

$$\frac{R_0}{2}\left[1+\cos 2\omega l\sqrt{(LC)}\right]-j\frac{R_0}{2}\left[\sin 2\omega l\sqrt{(LC)}\right] \quad . \quad (349)$$

corresponding to a single *positive* voltage echo.
These impedance/frequency characteristics are illustrated by Fig. 118 in terms of their real and imaginary parts; the echo patterns have already been shown in Fig. 117 (b). The "period" of these sinusoidal shaped characteristics is $\omega=\pi/l\sqrt{(LC)}$, and it should be noted that the corresponding echo delay time is $2l\sqrt{(LC)}$, which is thus simply related to the "period."

In these extreme short- and open-circuit cases in which $R_T=0$ or $\infty$ the echoes have the same magnitude as the applied signal, the line being assumed lossless, but negative and positive respectively in sign. With intermediate values of $R_T$ the echoes are smaller in magnitude, depending on the ratio $(R_T-R_0)/(R_T+R_0)$, and of course are of zero amplitude in the matched case $R_T=R_0$.

Now let us return to the line mis-matched at both ends; equations 345 and 346 give the sending-end impedance of a length of line open-circuited at the sending end and short- or open-circuited at the far end. Fig. 117 (c) illustrates the latter case and 117 (d) shows the corresponding echo pattern—a series of constant-amplitude signals (pulses in this case) spaced at regular time intervals.

* See the set of characteristics given in the discussion in the I.E.E. paper, reference 10, which concerns the mis-match caused by insertion of loading coils.

Now since the line which shows a single echo has a sinusoidal shape of impedance/frequency characteristic we might expect that the characteristic of this line showing multiple echoes may be divided up into a number of sinusoidal components. The "periods" of these components should correspond to the echo times $2\sqrt{(LC)}$, $4\sqrt{(LC)}$, $6\sqrt{(LC)}$ . . . etc. Take the case of the line with the open-circuited end, illustrated by Figs. 117 (c) and (d); the impedance/frequency characteristic is given by 346, and this expression may be expanded:

$$Z_s(\omega) = -jR_0 \cot \omega l\sqrt{(LC)} = R_0 \frac{(\epsilon^{j\omega l\sqrt{(LC)}} + \epsilon^{-j\omega l\sqrt{(LC)}})}{(\epsilon^{j\omega l\sqrt{(LC)}} - \epsilon^{-j\omega l\sqrt{(LC)}})} \quad . \quad . \quad (350)$$

which may be written:

$$\begin{aligned} Z_s(\omega) &= R_0(1+\epsilon^{-2j\omega l\sqrt{(LC)}})(1-\epsilon^{-2j\omega l\sqrt{(LC)}})^{-1} \\ &= R_0(1+\epsilon^{-2j\omega l\sqrt{(LC)}})(1+\epsilon^{-2j\omega l\sqrt{(LC)}}+\epsilon^{-4j\omega l\sqrt{(LC)}}+\epsilon^{-6j\omega l\sqrt{(LC)}} \\ &\qquad + \ . \ . \ . \text{ etc.}) \\ &= R_0(1+2\epsilon^{-2j\omega l\sqrt{(LC)}}+2\epsilon^{-4j\omega l\sqrt{(LC)}}+2\epsilon^{-6j\omega l\sqrt{(LC)}}+ \ . \ . \ . \text{ etc.}) \end{aligned}$$

or, in terms of its real and imaginary parts:

$$\begin{aligned} Z_s(\omega) = R_0\{1 &+ 2\cos[2\omega l\sqrt{(LC)}] + 2\cos[4\omega l\sqrt{(LC)}] \\ &+ 2\cos[6\omega l\sqrt{(LC)}] + \ . \ . \ . \text{ etc.}\} \\ -jR_0\{&2\sin[2\omega l\sqrt{(LC)}] + 2\sin[4\omega l\sqrt{(LC)}] \\ &+ 2\sin[6\omega l\sqrt{(LC)}] + \ . \ . \ . \text{ etc.}\} \end{aligned} \quad . \quad . \quad . \quad . \quad . \quad . \quad (351)$$

We have thus divided up the impedance given by 346 into a series* of sinusoidal-shaped characteristics, each of the form shown in Fig. 118 (b), having "periods" equal to $\pi/l\sqrt{(LC)}$, $\pi/2l\sqrt{(LC)}$, $\pi/3l\sqrt{(LC)}$ . . . etc. We have, in effect, analysed the impedance/frequency characteristics 346 into its Fourier spectrum of harmonics —harmonics of a certain fundamental period of *angular frequency* $\pi/2l\sqrt{(LC)}$. Such a Fourier analysis is the inverse of the analysis of transients carried out in Chapter 2, where the fundamental period is a period of *time*.

* Mathematical Note: The series 351 is not convergent; however, we are not interested in reversing the process and summing this series to show it equals $-jR_0 \cot \omega l\sqrt{(LC)}$. The term $(1-\epsilon^{-2j\omega l\sqrt{(LC)}})^{-1}$ has been expanded by the Binomial Theorem though the modulus of $\epsilon^{-2j\omega l\sqrt{(LC)}}$ is unity. We have started with a physically impossible condition (that the line has *zero* resistance). If dissipation exists, however small, this becomes $\epsilon^{-2j\omega l\sqrt{(LC)}-r}$ (where $r$ is real), so that its modulus is $<1$ and the series 351 then converges.

In a similar manner we could have analysed the characteristic 345 for the short-circuited length of line, as the reader may try for himself.

Each of the component sinusoidal characteristics in 351 may be considered to set up a single echo, as though it were produced by a length of line matched at the sending end but open-circuited at the far end.

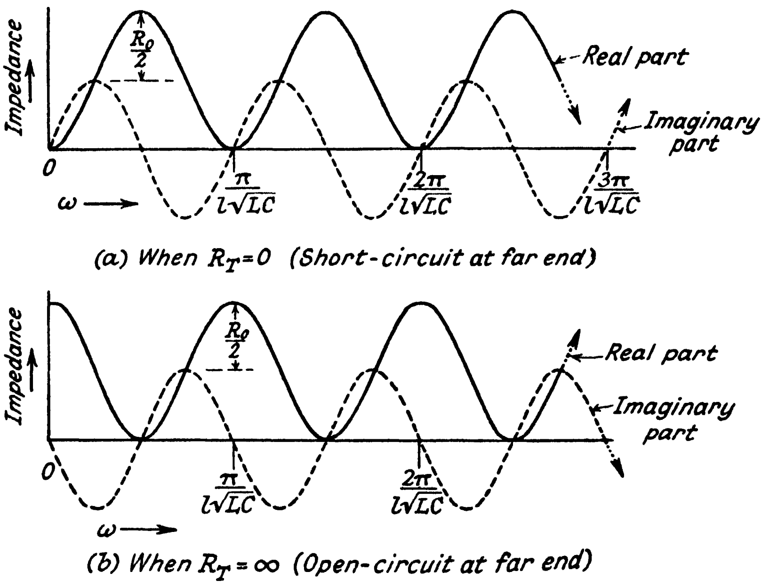



Fig. 118.—Impedance/Frequency Characteristics of the Line in Fig. 117 (a).

The echoes as given by the series 351 are all of the same magnitude. In practice successive echoes must become progressively smaller due to the inherent dissipation in the system (see footnote on previous page). If $N$ is any integer:

Echo delay time $= N \,.\, 2l\sqrt{(LC)}$

Length of reflecting line $= Nl$

"Period" of impedance/frequency characteristic $= \dfrac{\pi}{Nl\sqrt{(LC)}}$

$$\qquad \qquad \ldots \ldots \quad (352)$$

### 73. Reflections in artificial lines—the delay network

The type of characteristics discussed in the last section which produce perfect echoes of an applied signal are obtainable only with

a length of real line and they cannot be attained by an artificial line or by any other lumped network. If the length of line be replaced by an artificial line the resulting virtual echo will be a distorted version of the applied signal, since the network's frequency characteristics cannot have the exact sinusoidal shape (Fig. 118) which is associated with perfect echo production.

One of the types of artificial line commonly used reflectively is the constant-$K$ structure illustrated by Fig. 111. This structure is often used for delaying pulses or other signals in television and radar systems, and in such connection is often called the *delay network.* It is also used for generating nominally rectangular pulses by applying a voltage of step waveform to its input terminals ($ab$) and allowing the wave reflected from the short-circuited end ($cd$) to cancel out this applied wave at a time $\Delta t$ later, as shown in Fig. 111 (a). The delay $\Delta t$ is virtual only, since the network is a lumped one, and depends on the number of sections in the network and on the cut-off frequency. In practice the reflected step wave will be distorted by the imperfect frequency characteristics of the network and so the resulting pulse cannot be rectangular in shape. If, for a given delay time $\Delta t$, the number of sections in the network is increased the resulting pulse improves in shape, theoretically; as the number of sections approaches infinity the network becomes like a length of real, distortionless line.

The virtual delay time (and hence the generated pulse width) is best estimated from a knowledge of the phase-shift characteristic of this constant-$K$ network. Both the phase-shift component and the modulus of the frequency characteristics introduce distortion into the echo. These characteristics are most readily calculable for an infinite chain of sections [12, 6] and are of the general form illustrated by Fig. 119 (drawn as modulus and phase shift *per section*). Frequently in practice a few sections only are used and then the characteristics involve extremely laborious calculation [13, 14] ; they depart somewhat from those illustrated,* but a sufficiently good estimation of practical results may be made from these ideal curves, although any such estimation is necessarily somewhat rough-and-ready.

It is preferable to choose the cut-off angular frequency $\omega_c$ reasonably high, so that the spectrum of the applied signal covers the more linear lower end of the phase-shift characteristic (Fig. 119 (b)).

* See, for example, those plotted in Fig. 78 (a) for a single $\pi$ section.

It is then approximately correct to assume the slope of this characteristic is:

$$\frac{\phi(\omega)}{\omega}=\frac{\pi}{1{\cdot}3\omega_c} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (353)$$

corresponding to the straight dotted line $OP$ in the figure. This approximation then neglects phase-shift distortion. Equation 353 gives an effective delay time, per section, according to this rough empirical rule.*

If the total required echo time (go *and* return) is $\Delta t$, then the network will have to contain $N$ sections, where:

$$N=\frac{\Delta t}{2\pi}(1{\cdot}3\omega_c) \quad . \quad . \quad . \quad . \quad . \quad . \quad (354)$$

The number $N$ thus increases with $\omega_c$; but we have already seen that the greater this number of sections, for a fixed total delay, the more closely the network approaches the distortionless length of line. As $\omega_c$ is made greater the spectrum of the applied signal lies on a more linear part of the phase-shift characteristic, assuming that this spectrum is limited in extent. As $N$ is made very large the empirical constant 1·3 should be reduced to unity.

There are obvious practical limitations to the number of sections which may be used, and it is often a question of compromise between number of circuit elements and permissible distortion. An improvement of the linearity of the phase characteristic is possible by introducing mutual inductance [15, 12] between adjacent series elements, $L$, which converts the network into an $m$-derived filter structure, thereby increasing the frequency range over which the phase-shift characteristic is nearly linear. The best value of mutual inductance is given as $M=0{\cdot}1L$ approximately.

For the constant-$K$ delay network, Fig. 119, the cut-off angular frequency $\omega_c$ is:

$$\omega_c=2/\sqrt{(LC)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (355)$$

while the sending-end impedance (between terminals ($ab$)) is:

$$R_0\propto\sqrt{(L/C)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (356)$$

for both mid-series and mid-shunt driving conditions, these two impedances becoming equal at low-frequencies ($\omega \ll \omega_c$) to $R_0=\sqrt{(L/C)}$.

* For networks of a few sections the approximate formula (353) checks up fairly accurately with Carson's calculated values (see his plotted responses, reference 14).

No finite uniform constant-$K$ network (Fig. 111) can exactly simulate a length of reflecting line, whether it be short-circuited or open-circuited at its far end, since, for a given echo delay time, the number of sections in the network is an arbitrary choice. The more sections are used, the higher the cut-off frequency $\omega_c$ and the more closely the network approaches a length of real line. There are, however, certain *non-uniform* structures with an infinite number of elements which simulate short-circuited and open-circuited lines from the point of view of their sending-end impedances and consequently are suitable for echo production.

These sending-end impedances are given by 345 and 346 for short- and open-circuited lengths of line respectively. These may

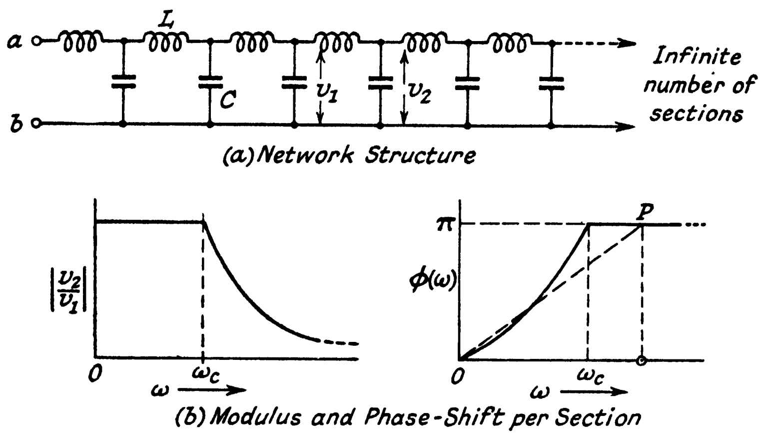



Fig. 119.—Constant-$K$ Delay Network. $\omega_C=2/\sqrt{LC}$.

be plotted, as curves of $Z_s(\omega)$ against $\omega$, showing how the impedances (pure reactances in these cases) vary with frequency; the curves are simple Tan and Cot functions and are illustrated in Fig. 120 (a). These impedance curves, which show an infinite series of resonances and anti-resonances, enable us to construct equivalent networks by using Foster's reactance theorem, as explained in Sec. 29; the equivalent networks must be composed entirely of reactances and have an identical series of resonances and anti-resonances. The possible types of network have been illustrated in Fig. 41, in which either a chain of anti-resonant circuits ((a) or (c)) are tuned to the anti-resonant frequencies on the reactance curve, or a chain of resonant circuits ((b) and (d)) are tuned to the resonant frequencies.

These critical frequencies are easily calculated, since they correspond to the frequencies of the natural modes of oscillation of the length of line. The mis-matched line becomes a tuned system, when driven in the steady state, by virtue of the interference between the forward-going and the reflected waves resulting in the setting up of the familiar standing-wave patterns [2, 4, 5] of voltage and current. A few of these patterns, for the lower frequency modes, are illustrated by Fig. 120 (b); the critical frequencies of resonance and anti-

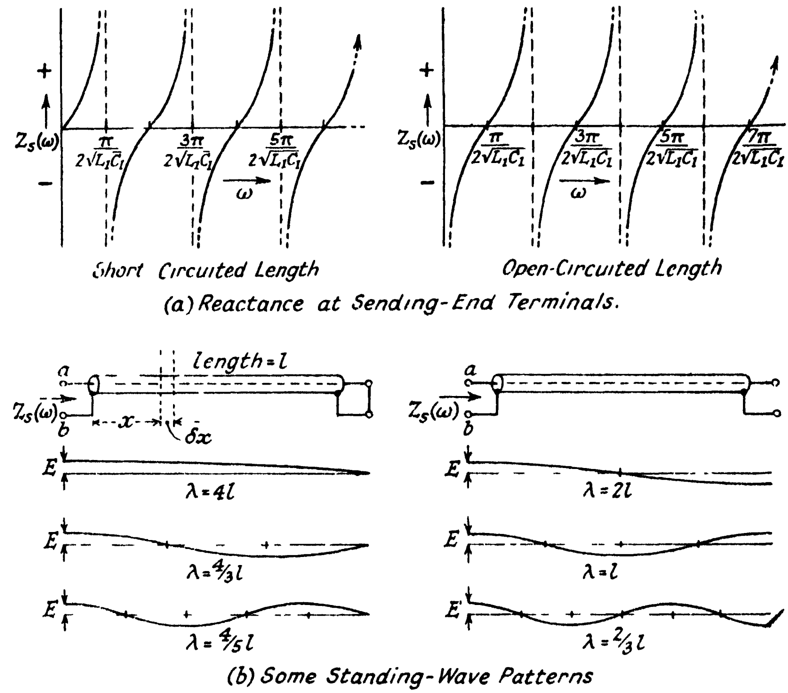



Fig. 120.—Sending-end Impedance, $Z_s$ ($\omega$), and Modes of Oscillation of a Tuned Line Length.

resonance are then expressible in terms of the number of quarter wavelengths into which the line length, $l$, divides. In this figure the *voltage* standing waves have been shown corresponding to anti-resonance, or infinite input impedance $Z_s(\omega)$. Two basic forms of continued structure may be used to simulate either the short- or the open-circuited line, and these forms are shown in Fig. 121. Either a chain of anti-resonant circuits tuned to the frequencies of infinite impedance, as in (a) and (c), or a chain of series resonant

circuits tuned to the frequencies of zero impedance, as in (b) and (d), may serve as equivalent impedances; in every case the chain must be considered to be infinitely long and hence cannot be reproduced physically without approximation.

Let $L_1$ and $C_1$ be the total inductance and capacity of the length of line, i.e.

$$\left.\begin{array}{l} L_1 = l \,.\, L \\ C_1 = l \,.\, C \end{array}\right\} \qquad \ldots\ldots\ldots \quad (357)$$

and the steady-state wave velocity along the line is $1/\sqrt{(LC)}$.

Now frequency = velocity/wavelength so that any particular critical frequency $\omega_n$ is given by:

$$\omega_n = \frac{2\pi}{\lambda_n} \cdot \frac{1}{\sqrt{(LC)}} = \frac{2\pi l}{\lambda_n \sqrt{(L_1 C_1)}} \qquad \ldots \quad (358)$$

in terms of $\lambda_n$, the particular standing-wave wavelength. It remains to express the wavelengths $\lambda_n$ in terms of $l$ for the various modes (see Fig. 120 (b)):

(1) *Short-Circuited Line.*

At frequencies of anti-resonance, i.e. voltage maxima across terminals ($ab$):

$$\lambda_n = 4l, \frac{4}{3}l, \frac{4}{5}l, \ldots \frac{4l}{(2n-1)} \ldots \text{etc.}$$

and at frequencies of resonance (current maxima):

$$\lambda_n = \infty, 2l, l, \frac{2}{3}l, \ldots \frac{2}{n}l \ldots \text{etc.}$$

(2) *Open-Circuited Line.*

At frequencies of anti-resonance:

$$\lambda_n = \infty, 2l, l, \frac{2}{3}l, \ldots \frac{2}{n}l \ldots \text{etc.}$$

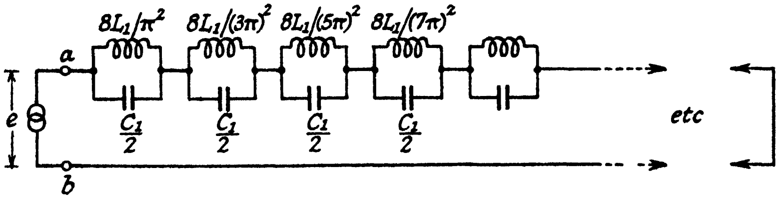
and at frequencies of resonance:

$$\lambda_n = 4l, \frac{4}{3}l, \frac{4}{5}l \ldots \frac{4l}{(2n-1)} \ldots \text{etc.}$$

Then substitution of these values for $\lambda_n$ in equation 358 gives the various anti-resonant and resonant frequencies to which the various tuned circuits in Fig. 121 must be adjusted, to simulate the input impedance of the line length.

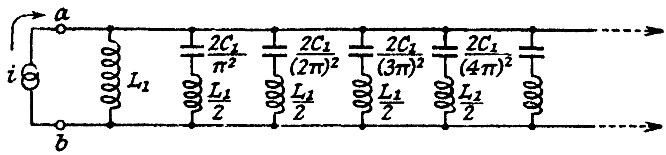
This has given us the tuning frequencies, but not the required inductance/capacity ratios. These can be found by considering the

energy stored in the line and in its equivalent network, both inductively and capacitively. Before this is done, it must be pointed out that the equivalent networks in Fig. 121 (a) and (c) can only be considered to be driven by a generator of known E.M.F., $e$, the



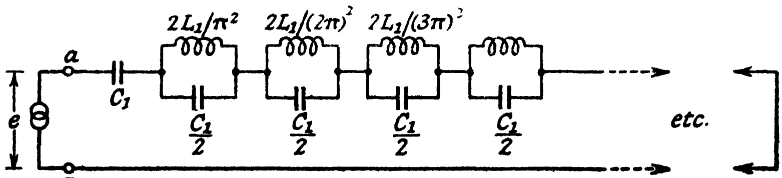

(a) Equivalent to Short-circuited Line.

—— Anti-resonant frequencies, $\omega = \frac{\pi}{2\sqrt{L_1C_1}}, \frac{3\pi}{2\sqrt{L_1C_1}}, \ldots \frac{(2n-1)\pi}{2\sqrt{L_1C_1}} \ldots$



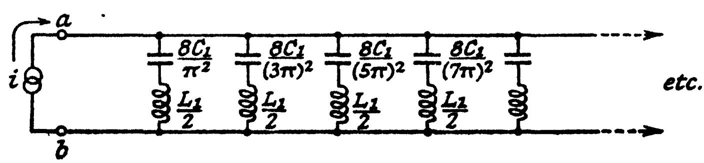

(b) Equivalent to Short-circuited Line.

—— Resonant frequencies, $\omega = 0, \frac{\pi}{\sqrt{L_1C_1}}, \frac{2\pi}{\sqrt{L_1C_1}}, \ldots \frac{n\pi}{\sqrt{L_1C_1}} \ldots$




(c) Equivalent to Open-circuited Line.

—— Anti-resonant frequencies, $\omega = 0, \frac{\pi}{\sqrt{L_1C_1}}, \frac{2\pi}{\sqrt{L_1C_1}}, \ldots \frac{n\pi}{\sqrt{L_1C_1}} \ldots$




(d) Equivalent to Open-circuited Line.

—— Resonant frequencies, $\omega = \frac{\pi}{2\sqrt{L_1C_1}}, \frac{3\pi}{2\sqrt{L_1C_1}}, \frac{5\pi}{2\sqrt{L_1C_1}}, \ldots \frac{(2n-1)\pi}{2\sqrt{L_1C_1}} \ldots$

Fig. 121.—Networks Equivalent to Short- or Open-circuited Line Lengths.

current entering the terminals (*ab*), at any frequency, being unspecified. These networks consist only of anti-resonant circuits and when the frequency of the driving E.M.F. coincides with any one of these anti-resonant frequencies, the whole of the E.M.F., *e*, appears across that particular anti-resonant circuit and the *whole* of the energy is then stored in that one circuit. Similarly the networks in Figs. 121 (b) and (d) can only be considered to be driven by a known *current*, *i*, since as each branch circuit resonates it short-circuits the line and the whole of the input current flows through it, storing the entire energy in it.

For brevity let us take only the case of the short-circuited line and its equivalent network (a) in Fig. 121. Let the frequency of the applied generator be such that the $n^{th}$ tuned circuit is anti-resonant, thereby containing all the energy in the network. The voltage distribution, $v_x$, along the equivalent line length is sinusoidal and in the $n^{th}$ anti-resonant mode it may be written (see Fig. 120 (b)):

$$v_x = E\cos\frac{2\pi x}{\lambda_n} = E\cos\frac{2\pi x(2n-1)}{4l} \quad . \quad . \quad (359)$$

The energy in any element of line $\delta x$ is $\frac{1}{2}C\delta x \, . \, v_x^2$, and the energy in the whole line length is therefore:

$$\text{Energy} = \tfrac{1}{2}\int_0^l CE^2\cos^2\frac{2\pi x(2n-1)}{4l}\,dx \quad . \quad . \quad (360)$$

$$= \tfrac{1}{4}Cl \, . \, E^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (361)$$

$$= \tfrac{1}{4}C_1E^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (362)$$

which is independent of *n*, the number of quarter wavelengths into which the line divides, and so independent of the frequency. Thus every capacity element in the equivalent network, Fig. 121 (a), must have the same value $C_1/2$ and contain the same capacitive energy at resonance, $\frac{1}{2}(C_1/2)E^2$. The values of the inductance elements then follow from the anti-resonant frequencies of the various tuned circuits.

If a current-driven equivalent circuit has been chosen instead, for example as in Fig. 121 (b), it would have been found that the energy at *every resonant* mode is $\frac{1}{4}L_1I^2$, so that each inductance element in this circuit must have the value $L_1/2$. Similarly for the open-circuited line and its two equivalent circuits, Figs. 121 (c) and (d), the element values all being marked in these diagrams.

These equivalent networks each have an infinite number of tuned circuits, but for practical purposes a finite number, and comparatively few, need be used since the elements converge rapidly to either open- or short-circuit values. They may be used as alternatives to the constant-$K$ delay network for echo production and pulse formation.

It should be observed that the sending-end impedances, 345 and 346, are reciprocal in form; the lumped networks representing them are therefore duals and, in Fig. 121, the network (d) could be formed by constructing the dual of (a) and similarly (c) from (b), according to the simple rules outlined in Sec. 28. The product of the sending-end impedances, 345 and 346, is $R_0^2=L/C=L_1/C_1$ giving the ratio of any two dual elements.

Thus the actual pieces of line, equal to one another in length, but short- and open-circuited respectively, are examples of extended networks which are duals (Fig. 120 (b)).

It is possible to arrive at the equivalent circuits of Fig. 121 by purely mathematical argument, by expanding the Tan and Cot expressions, 345 and 346, giving the short- and open-circuited line impedances as a series of terms involving $L_1$ and $C_1$. We have used our semi-physical approach here as being more suited to our purpose, since it exhibits more clearly the relations between extended and lumped equivalent systems. Bartlett [11] has carried out such expansions, starting from expressions based on the impedance of a lossy line, and his equivalent lumped circuits are similar to those in Fig. 121, but included resistance. Resistance may be added to these circuits provided that the Q values of every resonant and anti-resonant arm or mesh are made identical. Thus every resonant arm may have a series resistance added and every anti-resonant mesh a shunt resistance; the resistance values may be such as to give any required constant Q value (see Sec. 32).

## 74. Transient response analysis by the method of paired echoes

Although the responses of many types of network to an applied transient signal may bear little resemblance to this signal, nevertheless the idea of echoes may be extended to networks other than artificial lines and a method of analysis may be evolved which depends on this idea.

In Fig. 66 we have shown a transient to be divided up into a series of adjacent pulses or alternatively into a series of superimposed step waves. Other types of waveform could be used for dividing

up the transient, but these two, the pulse and the step wave, are particularly simple. Now let us suppose this transient $f(t)$, in Fig. 66 (a), to be the response of some network to an applied step wave; then each of the elementary step waves into which we have divided $f(t)$ may be considered to be an echo of the applied step wave. These echoes are spaced apart by the time $\delta\tau$, and are of various amplitudes.

In this way we may consider the network to be replaced by a number of lengths of transmission line, the step-wave signal being applied to the input terminals of each line and the lengths of line successively increased to give delay times differing by $\delta\tau$. These

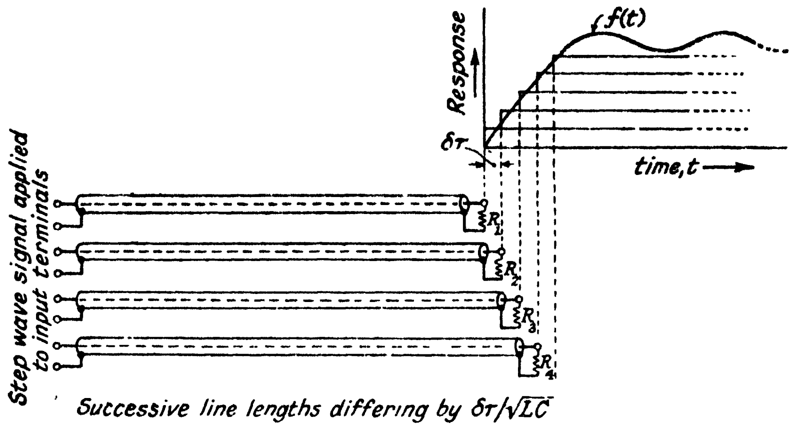



Fig. 122.—Synthesis of a Transient Response Wave by a Number of Lengths of Transmission Line.

(The first four lengths, only, are shown.)

lengths of line must be mis-matched at their receiving ends by different terminations in order that each may set up an echo of the appropriate amplitude. Such a system of lengths of line would then synthesise the response transient $f(t)$, approximately, by summation of the echoes received back at the input terminals. Fig. 122 illustrates this notion. The shorter the time $\delta\tau$ between echoes the more lengths of line are required and the closer the synthesised wave approaches $f(t)$, the small steps becoming smoothed out.

In the same way the signal divided into a series of pulse waves, Fig. 66 (b), could be synthesised by applying a pulse signal to a similar arrangement of line lengths.

In Fig. 122 the input terminals are not shown connected in

parallel, since in practice this would set up secondary echoes owing to the mis-match that would arise. This idea is purely a hypothesis, and the input step-wave signal must be considered to be applied instantaneously to the input terminals of all the lengths of line; similarly the echoes received back at these terminals must be considered to be added together without connection between the terminals. This could be achieved in practice by using a series of tetrode or pentode valves, one for driving each length of line.

A more practical way of synthesising a given transient signal waveform is by the use of a single length of line with tapping points placed at suitable intervals along its length. A step-wave voltage is applied to the sending-end terminals and arrives at the successive tapping points after various delay times. These delayed signals are then attenuated, to have the correct relative magnitudes, and combined by a series of valves. By observation of the resulting waveform on an oscillograph and adjustment of the delays and attenuations of the various component step waves, a transient response of a desired waveform may be built up.[16, 17]

In such a manner this line structure may be made to have the characteristics of a particular filter type by adjusting the response to the known transient response of the filter. A special case of interest lies in the synthesis of the response wave shown in Fig. 70 (the sine-integral form); the frequency characteristics of the structure then closely approach those of the idealised flat-topped filter with uniform phase shift.

Similar types of response-wave synthesis have been made using artificial lines.[15]

The representation, illustrated by Fig. 122, of a given network by a series of lengths of mis-matched line suggests that the frequency characteristics of the network may be analysed into a series of component characteristics each corresponding to those of a mis-matched line, i.e. the sinusoidal forms given by 347 and shown in Fig. 118 for the two extreme mis-match cases. Each set of sinusoidal component characteristics then results in an echo of the signal applied to the network's input terminals; the echo delay time for any one set is dependent on the "period" of the sinusoidal characteristics.

In this way, a Fourier analysis of the frequency characteristics of a network should enable us to determine the response of the network to an applied transient by the addition of an echo of the signal for each Fourier component in the analysis. This way of attacking

transient response problems has become known as the "method of paired echoes," [18, 8] and will be examined in greater detail.

Let $Z(j\omega)$ represent the characteristics of a network (2- or 4-terminal); a range of frequencies $W$ may be chosen, which adequately covers the spectrum of an applied signal. The curves of $Z(j\omega)$ may be written as a Fourier series with the range $W$ as the fundamental:

$$Z(j\omega)=r(\omega)+jx(\omega) \qquad . \quad . \quad . \quad . \quad (363)$$

where

$$\left.\begin{aligned} r(\omega)&=r_0+\sum_{n=1}^{\infty} r_c(n)\cos\frac{2\pi n\omega}{W}+r_s(n)\sin\frac{2\pi n\omega}{W} \\ x(\omega)&=x_0+\sum_{n=1}^{\infty} x_c(n)\cos\frac{2\pi n\omega}{W}+x_s(n)\sin\frac{2\pi n\omega}{W} \end{aligned}\right\} \quad . \quad (364)$$

The coefficients $r(n)$, $x(n)$ represent the amplitudes of the various harmonic curves, each of the form shown in Fig. 118, comprising the real and imaginary components of the characteristics $Z(j\omega)$; the suffixes $c$ and $s$ denote cosine and sine (even and odd) components, which depend on the chosen axis for the Fourier analysis. This axis may be chosen as the mid-band frequency for band-pass problems or as zero frequency (with conjugate characteristics, as in Fig. 33 (c)) for low-pass examples. In these latter cases, $r_s(n)=0$ and $x_c(n)=0$ owing to the symmetry.*

This Fourier analysis is theoretically possible in all cases, though it can be extremely laborious to carry out sometimes, depending upon the ultimate accuracy required.† Fig. 123 illustrates the band-pass and low-pass cases. At present, for simplicity, we shall take only the low-pass case or the symmetrical band-pass case (with mid-band transferred to zero frequency); usually also, $X_0=0$.

* It should be noted that $r_c(n)\cos 2\pi n\omega/W$ (the real cosine) and $x_s(n)\sin 2\pi n\omega/W$ (imaginary sine) component characteristics form the "in-phase" characteristics $Y_x(\omega)$ that were discussed in Chapter 7 (Asymmetric Sideband Channels). Similarly, $r_s(n)\sin 2\pi n\omega/W$ (the real sine) and $x_c(n)\cos 2\pi n\omega/W$ (the imaginary cosine) components, which can only arise in the case of asymmetric band-pass characteristics, form the "quadrature" or distorting characteristics $Y_y(\omega)$. We are suggesting here that once $Y_x(\omega)$ and $Y_y(\omega)$ have been calculated, according to Chapter 7, the actual wave distortion set up by such an asymmetric sideband channel may be calculated by this method of echoes.

† See the various references to harmonic analysis, including experimental methods, at the end of Chapter 2.

If now $I\cos\omega t$ is *one* component of a complex signal applied to the input terminals of the network, then the response wave $v(t)$ at the output terminals is, from 364:

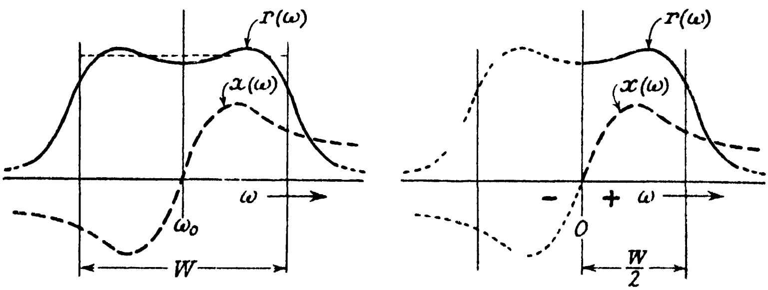
$$v(t)=r_0 I\cos\omega t+I\sum_{n=1}^{\infty} r_c(n)\cos\omega t\cos\frac{2\pi n\omega}{W}$$

$$+I\sum_{n=1}^{\infty} x_s(n)\sin\omega t\sin\frac{2\pi n\omega}{W}$$

since we are at present assuming $r_s(n)$ and $x_c(n)=0$. This may be written:

$$v(t)=r_0 I\cos\omega t+\frac{I}{2}\sum_{n=1}^{\infty} r_c(n)\cos\left(\overline{\omega t+\frac{2\pi n}{W}}\right)+r_c(n)\cos\left(\overline{\omega t-\frac{2\pi n}{W}}\right)$$

$$+\frac{I}{2}\sum_{n=1}^{\infty} x_s(n)\cos\left(\overline{\omega t-\frac{2\pi n}{W}}\right)-x_s(n)\cos\left(\overline{\omega t+\frac{2\pi n}{W}}\right) \quad . \quad . \quad (365)$$




(a) Band pass (b) Low-pass (Conjugate Characteristics)

FIG. 123.– Illustrating a Frequency Range, $W$, for Fourier Analysis of a Set of Characteristics.

The response signal corresponding to each value of $n$ consists of a component of amplitude $r_0 I$, corresponding to the applied wave $I\cos\omega t$, and two pairs of "echoes" of this wave displaced by time intervals $\pm 2\pi n/W$ about $t=0$, having amplitudes $r_c(n)\,.\,I/2$ and $x_s(n)\,.\,I/2$. Notice that the echoes given by the real and the imaginary parts of the network characteristic, $r_c(n)$ and $x_s(n)$ respectively, are of like sign, thereby adding, on the $t>0$ side, but are of opposite sign, thereby cancelling, on the $t<0$ side.

Physical reasoning demands that there can be no finite response before $t=0$, so that the sum of the echoes for $t<0$ must be zero.

This gives the relation $r_c(n)=x_s(n)$, which we have already seen to be true for the characteristics of a length of line; such characteristics are pure sinusoidal curves (Fig. 118) and the real and imaginary parts (given by equations 348 and 349) are of equal amplitude $R_0/2$. The resultant response wave, given by equation 365, then consists of the component of amplitude $r_0I$ and a series of echoes delayed by times $2\pi n/W$.

The delay time, $2\pi n/W$, is independent of frequency, but depends

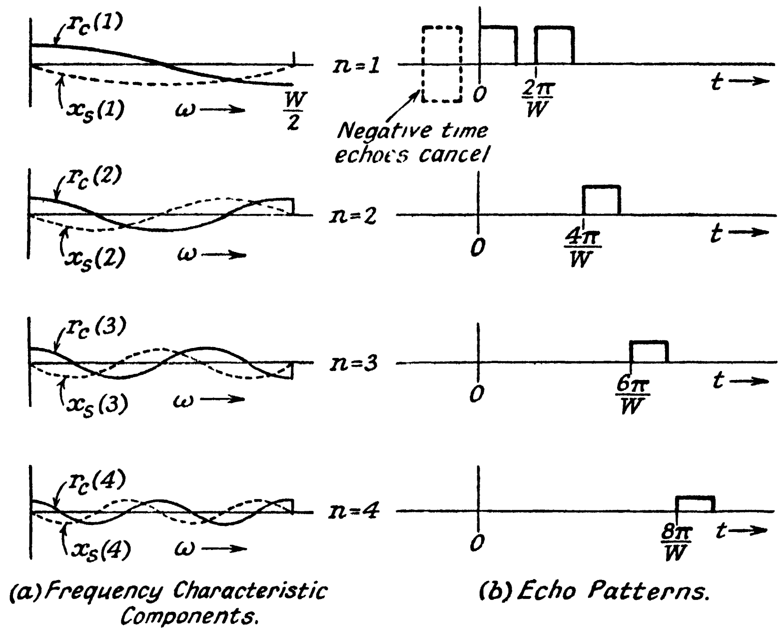



Fig. 124.—Fourier Components of a Network's Frequency Characteristics and the Corresponding Echoes Produced.

only on $n/W$, which is the period of a particular sinusoidal component of the Fourier series into which the network characteristics have been divided. Thus these pairs of echoes will appear when a signal of *any* waveform is applied to the input terminals, and not only in the steady-state case represented by equation 365. Fig. 124 shows a number of sets of sinusoidal components, corresponding to $n=1, 2, 3$ . . ., etc., into which we may imagine we have divided the characteristics of some network,* together with the individual

* $x_s(n)$ happens to be negative in these diagrams.

echo responses; the responses for $t<0$ must add up to zero. For this illustration a rectangular pulse wave has been assumed for the signal applied to the network's input terminals, which is shown together with the first echo, $n=1$. The echoes are separated here, but they may overlap in many cases. The echo amplitudes depend upon the magnitudes of the individual Fourier component characteristics, $r_c(n)$ and $x_s(n)$, and they are all *positive* echoes in this diagram.

Each of these sets of component characteristics is similar to that for the open-circuited line, Fig. 118 (b), which sets up a single positive echo; sometimes a network's characteristics may be analysed into a series of components like those of Fig. 118 (a), corresponding to a short-circuited line, in which case the echoes must be negative (i.e. inverted). Practical analysis in any given case decides which are to be used.

Modulus and phase-shift characteristics are more commonly used than are real and imaginary parts. They are of course related thus:

$$\left.\begin{aligned}&\text{Modulus} \quad |Z(\omega)| = \sqrt{[r(\omega)^2 + x(\omega)^2]}\\ &\text{Phase shift} \quad \theta(\omega) = \frac{x(\omega)}{r(\omega)}, \text{ if small}\end{aligned}\right\} \quad . \; . \; . \; . \; . \; . \; . \quad (366)$$

These may be used instead for interpreting amplitude and phase-shift distortion by this method of echoes,[8] though their direct use is approximate only, becoming more accurate as $\theta(\omega)$ approaches a uniform constant slope, $\theta(\omega) \rightarrow \omega t_1$. This implies that the method is accurate, for use with modulus and phase-shift curves, only when the phase distortion is small.

As before, a Fourier series over a given range of frequencies $W$ may be used to represent the characteristics and (within the limits of small phase distortion) a wavy modulus and phase-shift curve of purely sinusoidal shape corresponds to the setting up of a single echo of any applied signal. Fig. 125 (a) shows a typical set of filter characteristics, again symmetrical about either zero frequency or a mid-band $\omega_0$. An average phase slope $d\theta/d\omega = t_1$ has been drawn, as a dotted line, and this may be assumed to give a uniform delay of the signal and any echoes set up, so that it may be ignored in the assessment of distortion. Phase-shift distortion is caused by the departure of the phase-shift characteristic from this uniform slope, $\Delta\theta(\omega)$.

A working range of frequency $W$ may be chosen and these characteristics split up into a series of Fourier components having periods

$W$, $W/2$, $W/3$ . . . $W/n$, etc., Fig. 125 (b) shows one such set (the $n^{th}$ harmonic of $W$); the upper curve shows the sinusoidal modulus, $|Z(n)|$, and the lower curve the sinusoidal phase-shift departure from linearity, $\Delta\theta(n)$. These curves may be represented by the vector diagram shown, in a manner strictly analogous to that used for representing a sinusoidal waveform (see the vector diagrams in

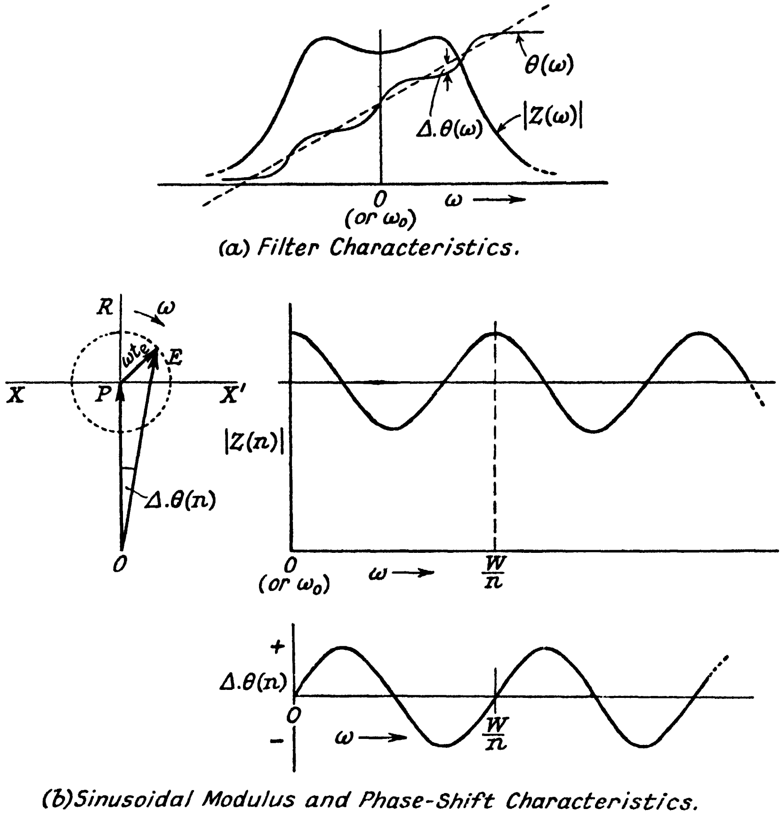



Fig. 125.—Echo Distortion from Modulus and Phase-shift Characteristics.

Chapter 2), though in the present case the vectors must be considered to rotate with *frequency* and not time.

Thus suppose $OP$ represents one steady-state component of an applied signal, $i = I\cos\omega t$, while $PE$ represents a single echo. If this echo is delayed in time by $t_e$ the angle between the vectors must be $\omega t_e$. The lengths $PE$ and $OP$ are proportional to the magnitudes

of the echo and of the applied signal. If $t_e$ is a constant echo delay time for every component of the signal (which may be a pulse or any transient), the phase angle between any component of frequency $\omega$ in the echo and the corresponding component in the signal must be $\omega t_e$—proportional to $\omega$. Thus the vector $PE$ rotates around the point $P$ as $\omega$ varies.

The vector $OE$ represents the resultant component of any frequency $\omega$, being the sum of the signal and echo. Its length varies almost sinusoidally, thereby tracing out the sinusoidal modulus characteristic, provided that $PE \ll OP$, that is, the echo/signal ratio is small. Similarly the phase angle $\underline{|POE}$ varies nearly sinusoidally and gives the sinusoidal phase-shift characteristic $\Delta\theta(n)$.

One period of the sinusoidal characteristics corresponds to a complete rotation of $PE$, or $\omega t_e = 2\pi$. Thus:

$$t_e = \frac{2\pi n}{W} \quad \ldots \ldots \quad (367)$$

gives the echo delay time.

The projections of the resultant vector $OE$ on to the real axis $OR$ gives the real component of the characteristics, which we have called $r_c(n)$, and the projection on to the imaginary axis $XPX'$ gives the imaginary component, $x_s(n)$:

$$\left.\begin{aligned} r_c(n) &= |OP| + |PE| \cos \omega t_e \\ x_s(n) &= |PE| \sin \omega t_e \end{aligned}\right\} \quad \ldots \quad (368)$$

It is possible to split up each vector in this diagram into two conjugate vectors; this would tend to confuse the diagram, but would show that the amplitude distortion alone (as represented by the sinusoidal modulus curve) requires two conjugate vectors of like sign, while the phase-shift distortion alone (sinusoidal phase-shift curve) requires two conjugate vectors of opposite sign. These represent the echoes spaced from the signal by times $\pm t_e$, and those echoes appearing before the signal must be equal and opposite. The vector diagram in Fig. 125 (b) gives this condition if $PE \ll OP$. If the modulus characteristic be drawn as a *ratio*, $PE/OP$, its variations must be equal in amplitude to those of the phase-shift characteristic, since $\underline{|POE} = |PE|/|OP|$ radians. Thus in terms of *echoes* the contributions to distortion (if small) by non-uniform modulus and phase-shift characteristics are equal; this compares to the equality of the effects of the real and imaginary parts, $r_c(n) = x_s(n)$, which we have already seen.

Such analysis is often very laborious, but cases arise in which the

distortion of a signal, in terms of echoes of the signal, may be estimated by inspection. For example, the characteristics shown in Fig. 123 have a pronounced sinusoidal "ripple" over the pass-band, with a period $W/2$ and magnitude $P$ per cent. of the mean response indicated by the dotted line. Thus the output signal from this network may be expected to consist of an echo added to the original signal, delayed by a time $4\pi/W$ and amplitude $P$ per cent. of the original signal. Characteristics of this particular form, for example, arise with a common type of intermediate-frequency filter, consisting of two tuned circuits coupled inductively (e.g. the circuit in Fig. 31 (b)).

However, this single echo cannot represent the total signal distortion, particularly if the signal spectrum reaches beyond cut-off. In such cases the characteristics may be considered to have the idealised rectangular or trapezium forms, as in Fig. 71, the distorted responses of which are readily calculable and have been plotted on this same figure (for an applied step wave). The sinusoidal component, of magnitude $P$ per cent., may then be added to this idealised characteristic and the practical characteristics thereby simulated more closely.

It must be emphasised that either the real part or the imaginary part of the characteristics needs to be known, but not necessarily both, when assessing echo distortion, since the echo patterns given by each part must be of identical form but of opposite sign for $t<0$, thereby cancelling out, and equal and of like sign for $t>0$. This point is examined more closely in the next section, when we come to express the network characteristics as a Fourier integral and not as a series; it will be appreciated that any series representation, such as we have used, is necessarily approximate since the characteristics are continuous curves extending to infinite frequency. Discrete, separated echo patterns as shown in Fig. 124 can only arise in real lengths of transmission line, and not in any lumped network, though great practical use may be made of a series representation when dealing with certain structures, such as delay networks and artificial lines, which can show pronounced echo effects.

## 75. Response of lumped networks in terms of pulse echoes

Instead of limiting the frequency range $W$ (see Fig. 123), we may allow this to extend indefinitely. Then instead of dividing up the frequency characteristics $Z(j\omega)=r(\omega)+jx(\omega)$ into a Fourier series of component characteristics having periods which are harmonic to

$1/W$, it will be necessary to have a continuous range of such characteristics. The corresponding echo pattern will then not consist of a series of discrete signal echoes spaced by time intervals $2\pi/W$ (Fig. 124), but will be continuous, having an infinitesimal time interval between successive echoes.

This, of course, implies that instead of analysing the characteristics as a Fourier series over the range $W$, a Fourier integral representation must be used. If the characteristics be expressed as a Fourier series, they may be imagined as being repeated indefinitely with the period $W$, that part lying outside the range $W$ being of no significance. The terms in the Fourier series then represent the

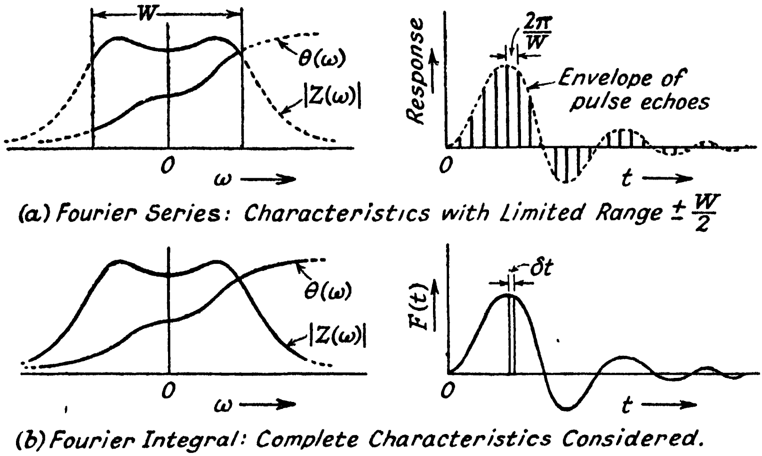



Fig. 126.—Response of Network $|Z(\omega)|$, $\theta(\omega)$ to an Infinitesimal Pulse: the Fourier Series and Integral Representation.

amplitudes of the echoes in the response to any applied signal. If this signal is an indefinitely short pulse the echo pattern will consist of a series of such pulses, spaced by time intervals $2\pi/W$, which may be considered as the *time* spectrum (see Fig. 126 (a)) of the characteristics. If $W \rightarrow \infty$, the interval $2\pi/W$ approaches zero, and the pulse echoes become adjacent, forming a smooth transient response $F(t)$ (Fig. 126 (b)); this response is the Fourier integral of the frequency characteristics.

Now the question arises: what physical significance can we attach to the Fourier integral of a set of frequency characteristics? This question has already been answered in Sec. 43, in which it was shown that this function is the response transient of the network

when an infinitesimally short pulse is applied to its input terminals. This may be shown as follows:

Such a pulse has a uniform spectrum, every component having the same magnitude (tending to zero as the pulse duration tends to zero). If this pulse is applied to the network of impedance $Z(j\omega)$ the response transient will have the spectrum $Z(j\omega)$, i.e. identical with the network characteristics. Thus this transient is the Fourier integral of $Z(j\omega)$; conversely, $Z(j\omega)$ is the Fourier integral of the pulse response transient—the two are Fourier Transforms.*

Thus the response wave $F(t)$ in Fig. 126 (b) may be considered to consist of an infinite number of adjacent pulses, of duration $\delta t$ (tending to zero), one of which is shown in this figure. Each consecutive pulse may then be considered as an echo of the pulse applied to the network. If any other signal, say $f(t)$, had been applied instead, the response transient would be obtained by

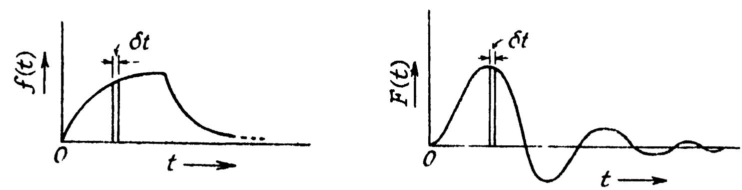



Fig. 127.—Transient Response by Pulse Echoes.

substituting echoes of $f(t)$ for every one of these pulse echoes, having corresponding amplitudes. This can readily be pictured in the series case, illustrated by Fig. 126 (a), but in the continuous integral response this continuous series of echoes can only be expressed operationally. Let $F(t)$ be the pulse response wave of the network having the impedance $Z(j\omega)$, shown in Fig. 126 (b); then $F(t)$ and $Z(j\omega)$ are Fourier Transforms. The operation of substituting echoes of the signal $f(t)$ for every adjacent pulse which comprises $F(t)$ (i.e. every ordinate) may be written:

$$\text{Response to applied signal } f(t) = f(t) \times F(t) \quad . \quad . \quad (369)$$

the sign $\times$ being used to indicate the particular operation defined. Fig. 127 illustrates this further; $f(t)$ is the applied signal and $F(t)$ is the network's response to an infinitesimal pulse, which itself is

* The mutually reciprocal relation between a function and its Fourier integral has been dealt with in Sec. 22. The present argument gives a particularly simple physical interpretation.

divided up into adjacent pulses, of duration $\delta t$ (tending to zero), as shown.

The operation represented by equation 369 may be applied in the reverse way. Thus we may divide the signal $f(t)$ into an infinite number of adjacent pulses, of duration $\delta t$, as shown in the figure. Then the response of the network to the signal $f(t)$ is obtained by adding responses of these pulses, each of which has the form $F(t)$. We may regard this as an addition of echoes; for every pulse into which we divide $f(t)$ we substitute an echo of $F(t)$, each echo having an amplitude proportional to its parent pulse. This operation may be represented by:

$$\text{Response to applied signal } f(t) = F(t) \times f(t) \quad . \quad . \quad (370)$$

From this it appears that the method of echoes, as a method of transient response analysis, follows directly from the Superposition Theorem, and consists merely of an addition of a number of pulse responses; an infinite number for an exact solution, but a finite series for most practical calculations.

It is important to appreciate the elementary fact that a set of frequency characteristics, such as those in Fig. 126, may represent either the steady-state behaviour of some physical network or it may be the spectrum of some transient. This transient will then be the pulse response (i.e. Fourier Transform) of the network. The behaviour of a network may be specified in two ways: by the steady-state characteristics or by the response to a transient of any known form; either is sufficient for a complete calculation or assessment of the response of the network to any signal, using the Superposition Theorem. The simplest transient is the infinitesimal pulse, but a more commonly used waveform is the step wave or Heaviside Unit Function. The pulse may be regarded as the differential coefficient* of the step wave, and the responses of a network to these two types of signal will themselves be related differentially (see Sec. 44).

The relation which we have shown between the steady-state and the pulse responses of a network is a physical example of the general idea of a Fourier Transform. Both responses represent the same thing (the behaviour of the network), but translated in terms of *time* or of *frequency*.[19] We may call the steady-state response $Z(j\omega)$ and the pulse response $F(t)$ the frequency characteristics and the time characteristics respectively.

* As regards waveform, but not absolute magnitude.

## 76. The inverse relations between frequency and time characteristics

Application of the inverse relation between a pair of Fourier Transforms to the frequency and time characteristics of a network provides some interesting comparisons. Briefly, the inverse relation (see Sec. 22, note (c)) between two Fourier Transforms implies that if $F(t)$ is the Fourier integral of $Z(j\omega)$, then $Z(j\omega)$ is itself the Fourier integral of $F(t)$. The variables $t$ and $f$ $(=\omega/2\pi)$ may be interchanged; note that $t$ has the dimensions of $1/f$. A simple example can make the significance of this more apparent: Fig. 128 (a) shows a simple sinusoidal wave—a pure A.C. wave—and its conventional representation by two conjugate vectors; also the spectrum is shown, consisting of two conjugate components of frequencies $\pm 1/T$ (where $T$=period of wave). Fig. 128 (b) shows the same set of diagrams, but with frequency $f$ and time $t$ everywhere interchanged. The "wave" is now a sinusoidal frequency spectrum, continuous between $f=\pm\infty$, and is represented by two conjugate vectors rotating with *frequency* (as previously used in Fig. 125). The "spectrum" of this wave (that is the Fourier series or integral), consists again of two components spaced by the *time* intervals $\pm 1/F$ (where $F$ is the period of the wave). The analogy is exact.

This "spectrum" consists simply of two pulses, and we could regard the spectrum of the A.C. wave in (a) as being two "pulses" in frequency. It is easy to show that two such pulses, spaced $\pm 1/F$, have a sinusoidal spectrum if we assume that a single pulse, of infinitesimal duration, consists of a continuous and uniform spectrum of cosine terms of the form $\Delta E \,.\, \cos\omega t$, where $\Delta E$ is indefinitely small. Then two pulses, spaced $\pm 1/F$, have a spectrum of terms of the form:

$$\Delta E\left\{\cos\omega\left(t-\frac{1}{F}\right)+\cos\omega\left(t+\frac{1}{F}\right)\right\}$$

$$=2\Delta E\left\{\cos\frac{\omega}{F}\right\}\,.\,\cos\omega t \quad . \quad . \quad . \quad . \quad . \quad . \qquad (371)$$

which is a spectrum of cosine components, varying in amplitude with frequency as $2\Delta E(\cos\omega/F)$, as shown in Fig. 128 (b). It must be remembered that there is no *magnitude* correlation between these diagrams, (a) and (b), but only one of geometric shape.

In diagram (a) the sinusoidal A.C. wave may be divided up into a series of adjacent pulses of duration $\delta t$ ($\rightarrow$ zero), whereas in (b)

the sinusoidal spectrum may be divided into similar vertical sections, but each one represents a sinusoidal component, with an amplitude given by equation 371. The pulse and the sinusidal curve are natural counterparts on a time scale and a frequency scale.

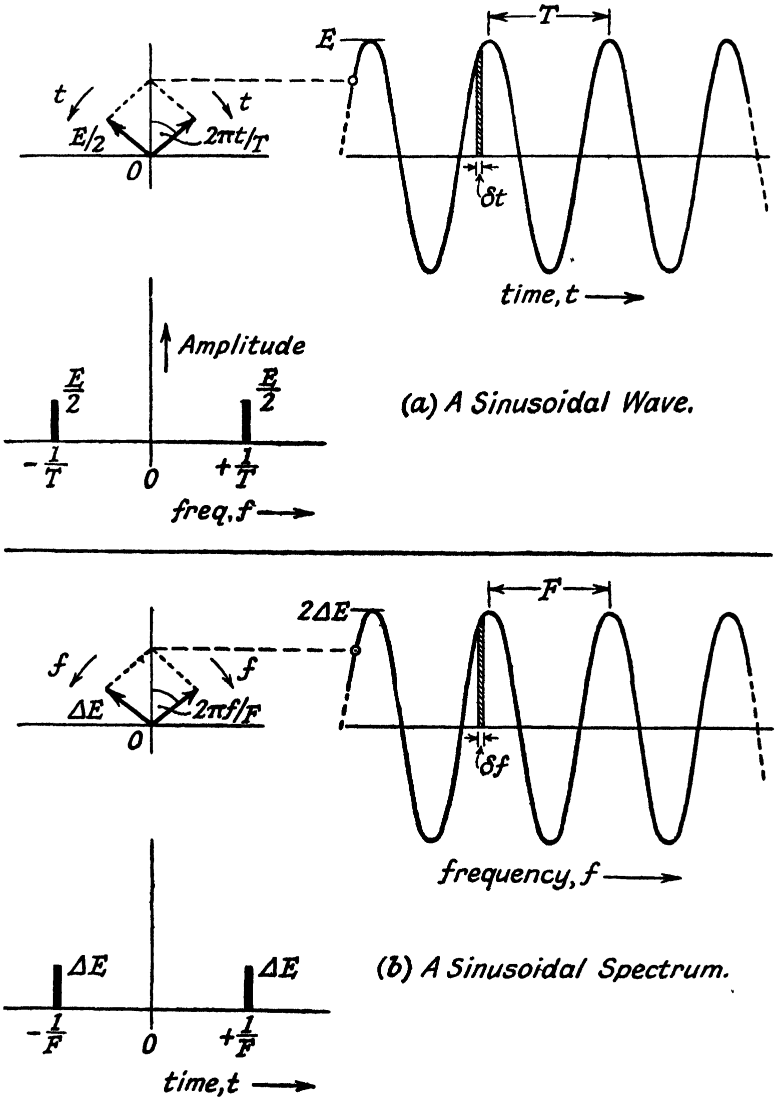



Fig. 128.—Comparison Between Functions of Frequency and of Time.

Another interesting case is provided by the analogy to the theorem concerning the transfer of a carrier frequency to any other frequency, including zero (see Sec. 16). Briefly, this theorem states that the spectrum of an amplitude modulated wave (having any envelope) is identical with the spectrum of its own envelope, except

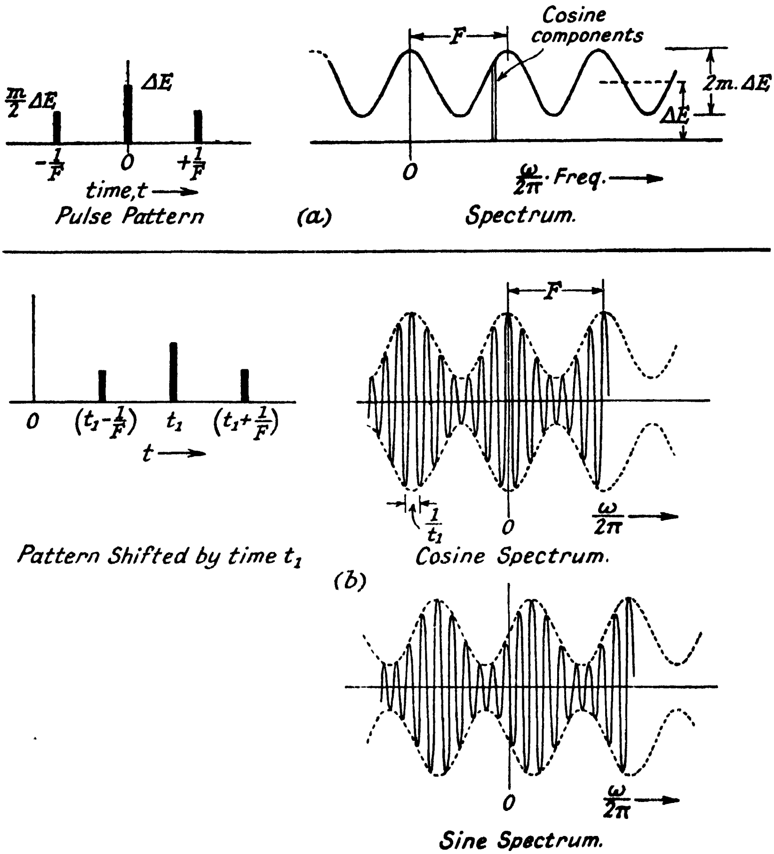



Fig. 129.—Transfer of a Pulse Pattern on the Time Scale. (Analogy to "Transfer of Carrier" Theorem.)

that the later is symmetrical about zero frequency while the former is symmetrical about the carrier frequency. Fig. 18 illustrates such "transferred" spectra. Let us consider the interchange of time and frequency in this theorem.

Fig. 129 (a) shows three pulses, symmetrical about zero time ($t=0$).

The spectrum of this pulse group consists of a continuous series of cosine terms, a term of any frequency $\omega$ being:

$$\Delta E \,.\, \left\{\cos \omega t + \frac{m}{2} \cos \omega\left(t - \frac{1}{F}\right) + \frac{m}{2} \cos \omega\left(t + \frac{1}{F}\right)\right\}$$

$$= \Delta E \left\{1 + m \cos \frac{\omega}{F}\right\} . \cos \omega t \quad . \quad . \quad . \quad (372)$$

This spectrum is sinusoidal in shape with a steady "D.C. component" of amplitude $\Delta E$. This is analogous to an A.C. wave added to a D.C. component (Fig. 19 (a)) and is shown plotted in Fig. 129 (a); every ordinate represents a cosine component.

Now, transferring the time origin to $t = t_1$ gives us the pulse group shown in (b), which has a spectrum of the form:

$$\Delta E \left\{1 + m \cos \frac{\omega}{F}\right\} . \cos \omega(t - t_1) \quad . \quad . \quad . \quad (373)$$

by shifting the time origin in equation 372. But this may be written:

$$\Delta E \left[\left(1 + m \cos \frac{\omega}{F}\right) \cos \omega t_1\right] . \cos \omega t$$

$$+ \Delta E \,.\, \left[\left(1 + m \cos \frac{\omega}{F}\right) \sin \omega t_1\right] . \sin \omega t \quad . \quad (374)$$

which splits the spectrum up into sine and cosine component spectra. Each has the form of a "modulated wave" with a carrier period $1/t_1$ and an "envelope" period $F$. This is also illustrated in Fig. 129 (b).

This is very similar to the theorem concerning the transfer of the carrier frequency of a modulated wave. The same process may be demonstrated, starting with a skew-symmetrical pair of pulses about a central component, which merely alters the phase of the "envelope" of the spectra about zero frequency. Vector diagrams may be constructed, in the same manner that was used in Fig. 128, for these three pulses and their projections on a fixed axis generate the sinusoidal spectra that we have calculated. These vector diagrams would be similar to those representing a conventional amplitude-modulated wave (see Fig. 20), but with the variables of time and frequency interchanged.

## REFERENCES

1. E. C. Cork and J. L. Pawsey, "Long Feeders for Transmitting Wide Sidebands, with Reference to the Alexandra Palace Aerial-Feeder System," *J.I.E.E.*, Vol. 84, No. 508, p. 448, April 1939.

2. W. L. EVERITT, "Communication Engineering," McGraw-Hill, New York and London, 1937.
3. G. W. O. HOWE, "Reflections from an Unmatched Feeder," *Wireless Engineer* (Editorial), Vol. 20, No. 236, May 1943.
4. W. JACKSON, "High Frequency Transmission Lines," Methuen Monograph, London, 1945.
5. W. T. PALMER, "Outline Notes on Telephone Transmission Theory," Isaac Pitman, 1935.
6. A. T. STARR, "Electric Circuits and Wave Filters," Isaac Pitman, London, 1938, Chapter VII.
7. H. W. BODE and R. L. DIETZOLD, "Ideal Wave Filters," *Bell Syst. Tech. Jour.*, Vol. 14, p. 215, April 1935.
8. E. GUILLEMIN, "Communication Networks," John Wiley, New York (Chapman & Hall, London), Vol. II, 1935.
9. G. A. CAMPBELL, "Loaded Lines in Telephonic Transmission," *Phil. Mag.*, Vol. 5, March 1903.
10. Sir WILLIAM NOBLE, "The Long-Distance Telephone System of the United Kingdom," *J.I.E.E.*, Vol. 59, p. 389, 1921.
11. A. C. BARTLETT, "The Theory of Electrical Artificial Lines and Filters," Chapman & Hall, 1930, Chapter V.
12. T. E. SHEA, "Transmission Networks and Wave Filters," Van Nostrand, New York.
13. H. WHEELER and F. MURNAGHAN, "The Theory of Wave Filters containing a Finite Number of Sections," *Phil. Mag.*, (7), Vol. 6, p. 146, 1928.
14. J. R. CARSON, "Electric Circuit Theory and the Operational Calculus," *Bell Syst. Tech. Jour.*, Vol. 5, p. 50, Jan. 1926 (transient currents in successive sections of artificial line).
15. M. LEVY, "The Impulse Response of Electrical Networks; with Special Reference to the Use of Artificial Lines," *J.I.E.E.*, Vol. 90, Part III, No. 12, 1943.
16. H. KALLMAN, "Transversal Filters," *Proc. I.R.E.*, Vol. 28, p. 302, July 1940.
17. A. BLUMLEIN, British Patent 517516.
18. H. WHEELER, "The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes," *Proc. I.R.E.*, Vol. 27, p. 359, 1939.
19. E. C. CHERRY, "Pulse Response: A New Approach to A.C. Electric Network Theory and Measurement," *J.I.E.E.*, Vol. 92, Part III, No. 19, Sept. 1945.

The following are also of interest:

(A) F. F. ROBERTS, "New Methods for Locating Cable Faults, particularly on High Frequency Cables," *J.I.E.E.*, Vol. 93, Part III, Nov. 1946.

(B) T. SHAW and W. FONDILLER, "Development and Application of Loading for Telephone Circuits," *Trans. A.I.E.E.*, 1926.

(C) W. P. MASON, "A New Method for obtaining Transient Solutions of Electrical Networks," *Bell Syst. Tech. Jour.*, Vol. 8, p. 109, Jan. 1929 (simulation of single circuit elements by sections of artificial line).

## APPENDIX

## THE "PROBABILITY" FUNCTION RESPONSE CHARACTERISTIC

It has been stated in Sec. 53, Chapter 6, that the overall frequency characteristic of an $N$-stage tetrode amplifier using shunt resistance-capacity loads is equal to $|Z(\omega)|$, where:

$$|Z(\omega)| = |Z(0)| \,.\, [1+\omega^2C^2R^2]^{-N/2} \quad . \quad . \quad [(249)]$$

where $|Z(0)| = (gR)^N$, the value at zero frequency, as illustrated in Fig. 82 (a) for values of $N$ up to six stages. This figure has been plotted in conjugate form so as to serve also for the equivalent band-pass circuit, which uses shunt tuned loads, inductance-capacity-resistance in parallel. Furthermore it has been stated that this characteristic assumes the "probability" function form, given by equation 236, as $N$ tends towards infinity. This curve is also shown in Fig. 82 (a) by the dotted line. The proof of this is as follows:

In equation 249, above, we may keep the band-width constant at $\omega = \omega_1$, the frequency at which $|Z(\omega)| = |Z(0)|/\sqrt{2}$, which as we saw in Sec. 53 is accomplished by making $CR$ a function of $N$. The required function is given by equation 250; substituting this in 249, above, gives for the overall amplifier characteristic:

$$|Z(\omega)| = |Z(0)| \,.\, \left[1+\frac{\omega^2}{\omega_1{}^2}(2^{1/N}-1)\right]^{-N/2} \quad . \quad (A.1)$$

But
$$(2^{1/N}-1) \rightarrow \frac{(\log_\epsilon 2)}{N} \text{ as } N \rightarrow \infty \quad . \quad . \quad . \quad (A.2)$$

since, if we write $(2^{1/N}-1) = (\epsilon^{(1/N)\log_\epsilon 2}-1)$ and expand:

$$\epsilon^{(1/N)\log_\epsilon 2} = 1+\frac{\log_\epsilon 2}{N}+\frac{(\log_\epsilon 2)^2}{N^2\,.\,2!}+ \ . \ . \ . \text{ higher power of } \frac{1}{N}$$

$$\text{which} \rightarrow 1+\frac{\log_\epsilon 2}{N} \text{ as } N \rightarrow \infty$$

Substituting this for $2^{1/N}$ in equation A.1 gives (if we write $\log_\epsilon 2 = 0{\cdot}69$):

$$|Z(\omega)| = |Z(0)| \,.\, \left[1+\frac{2}{N}\left(\frac{\omega}{1{\cdot}7\omega_1}\right)^2\right]^{-N/2} \quad . \quad . \quad (A.3)$$

Now this is of exponential form, since we have by definition:

$$\epsilon^{x} = \lim_{N \to \infty} \left(1 - \frac{x}{N}\right)^{-N} \quad . \quad . \quad . \quad . \quad . \quad . \quad (A.4)$$

so that, as $N \to \infty$, equation A.3 becomes:

$$|Z(\omega)| = |Z(0)| \, . \, \epsilon^{-\left(\frac{\omega}{1 \cdot 7\omega_1}\right)^2} \quad . \quad . \quad . \quad (A.5)$$

which is the "probability" function characteristic, equation 236.

This analytic form of characteristic is very useful for making approximate calculations, as we have seen in Chapter 6, since although it represents the theoretical frequency response of an amplifier with an infinite number of identical stages, it is very nearly correct for comparatively few stages, as may be used in practice, since the expression A.1 is very rapidly convergent with $N$.

# INDEX

PRINTED AND BOUND IN ENGLAND BY
WILLIAM CLOWES AND SONS LIMITED
LONDON AND BECCLES